Detecting Sequential Patterns and Determining Their Reliability With Fallible Observers

Roger Bakeman
Georgia State University

Vicenç Quera
University of Barcelona

Duncan McArthur
Georgia State University

Byron F. Robinson
Emory University

On the basis of recent work by W. Gardner (1995), implications of fallible observers for observational research are discussed. Analysis shows that for identically fallible observers, values for kappa are lower when codes are few and their simple probabilities variable than when codes are many and roughly equiprobable; thus no one value of kappa can be regarded as universally acceptable. Additional analysis shows that fallible measurement degrades indices of sequential pattern more when codes are few and their simple probabilities variable. Finally, a simulation study establishes likely values for intraclass correlation reliabilities for sequential indices generated by various circumstances and suggests principled ways to select both lengths of sequences and acceptable levels of kappa for observational studies.

Investigators who systematically code behavior in an attempt to detect recurring patterns usually take considerable care to establish that observers viewing the same events agree as to how the behavior should be coded (Bakeman, in press). Later we discuss the merit of determining whether indices of sequential pattern are similar when sequences of events are coded by different observers but, in the first instance, most investigators regard it as sufficient to establish that observers agree when coding single events. For example, assume that a set of $K$ mutually exclusive and exhaustive codes is defined, designated $C_1$ through $C_K$, and that two observers have independently assigned one of these codes to each of $N$ sequential events. Typically, the codes the observers assign to each event are tallied in a $K \times K$ agreement matrix. Columns represent one observer, rows the other, both rows and columns are labeled $C_1$ through $C_K$, and Cohen's kappa (Cohen, 1960), a statistic that corrects for chance agreement, is computed and reported (Bakeman & Gottman, 1986, 1997).

When the agreement matrix represents judgments of two independent observers, kappa is regarded most cautiously as an index of interobserver agreement, and is so regarded here. Some investigators may regard kappa as an index of observer reliability—that is, as an index of the extent to which measurement error is absent from the data (Nunnally, 1978)—but to do so requires that the essentially untestable classical parallel test assumptions be met (Suen, 1988). Of course, if an observer was compared with a protocol known (or assumed) to represent the true state of affairs, then the agreement matrix and the kappa derived from it would represent observer reliability. However, whether regarded as interobserver agreement or observer reliability, investigators have traditionally calculated kappa with reference to single or simple events, not patterns of those events.

Investigators understand that observers are fallible, which is why the usual interobserver kappa is routinely reported. Such kappas are useful because, assuming that observers’ errors are random and independent of each other, they suggest a lower bound for observer reliability for simple events. After all, if one of the randomly fallible observers were replaced with...
an infallible one, one who knew the true state of affairs, then some misclassification errors would disappear from the agreement matrix and the value for kappa would increase. However, the inaccurate measurement of simple events produced by a fallible observer also affects sequential patterns, and kappa for simple events does not directly assess the reliability of measurement for such patterns, which usually are central to the substantive questions asked by observational researchers.

In sum, fallible coding has implications, and some may not be widely appreciated. Three implications are explored here. First, we consider how the usual interobserver kappa should be interpreted given fallible observers and whether values that are unacceptable in some circumstance might be acceptable in other circumstances. Second, on the basis of recent work by Gardner (1995), we consider whether detecting sequential patterns is more disrupted in some circumstances than others by fallible observers. Finally, we consider the effect of fallible observers when traditional reliability indices are derived, not for single events but for sequential patterns.

Fallible Observers and Interobserver Kappa

If an observer's coding of events was compared with a known, accurate standard, then kappa would represent reliability and diagonal elements of the agreement matrix would represent the observer's accuracy for individual codes. Usually, however, two fallible observers are compared. If both observers tended to make similar mistakes when coding events, then interobserver kappa could be higher than the kappa estimating reliability (i.e., comparing a fallible to an infallible observer), but if both observers' errors were random and independent, then interobserver kappa will represent a lower bound for reliability. Nonetheless, some ambiguity surrounds the magnitude of kappa; values may be significantly different from zero (Fleiss, Cohen, & Everitt, 1969) but not sufficiently large to satisfy an investigator's desire for accuracy.

Factors That Affect Kappa

To provide a basis for deciding when values of kappa are sufficiently large, we varied factors that affect values of kappa and then examined the values of kappa that resulted. Three factors were varied: the number of codes (i.e., $K$), the variability of their simple probabilities, and the accuracy with which the two observers code simple events. We let $K$ vary from 2 through 10 because such values are often encountered and any trends noted should extrapolate to larger values. The lower values may be encountered less frequently, but they often represent a worst case scenario and for that reason are useful. For simplicity, we report results for $K = 2, 3, 5$, and 10 because these seem sufficient to establish trends.

For each value of $K$, we defined three levels of simple probability variability: equiprobable, moderately variable, and highly variable (see Table 1). Let $\pi_i$ represent the true probabilities for the $K$ codes. For the equiprobable case, $\pi_i = 1/K$ for all $i$, 1 through $K$; such a case is rarely encountered but is included here as a baseline. For the moderately variable case, $\pi_i = 0.5/K$, $\pi_K = 1.5/K$, and other probabilities assumed graduated intermediate values as shown in Table 1. Similarly, for the highly variable case, $\pi_i = 0.25/K$, $\pi_K = 1.75/K$, and other probabilities again assumed graduated intermediate values. Rarely would probabilities be so neatly graduated in an actual investigation, but the differences between smallest and largest
test probabilities represented by our three cases should provide some guidance when investigators encounter a similar range of differences in their simple probabilities.

Central to the present discussion is the accuracy (or fallibility) of observers. Let \( p \) represent the accuracy with which an observer codes simple events; when required, \( \sigma \) represents accuracy for a second observer (\( p \) and \( \sigma \) represent observer reliability for simple events, but to avoid confusion with later discussion of reliability for patterns of events, we consistently refer to \( p \) and \( \sigma \) as observer accuracy). Assume that the accuracy of each observer is known and is summarized in a table of conditional probabilities. For the first observer

\[ P_{ij} \]

is the conditional probability that an event will be coded \( C_j \) given that the event really is \( C_i \) (thus \( P_{ij} \) refers to the \( i \)th row and \( j \)th column of the conditional probability matrix). If measurement were perfect (i.e., if the observer were infallible) then \( P_{ij} \) would equal 1 when \( i = j \) (e.g., \( C_1 \) events would always be coded \( C_1 \)) and \( P_{ij} \) would equal 0 when \( i \neq j \) (e.g., \( C_1 \) events would never be coded \( C_2, C_3 \), etc.). If measurement were imperfect, however, elements on the diagonal of the \( p \) matrix would be less than 1 and off-diagonal elements would be greater than 0; likewise for \( \sigma \) and the second observer. To avoid confusion, one should remember that \( p \) and \( \sigma \) are tables of conditional probabilities, in which the rows (but not the columns) must sum to 1, whereas the kappa agreement matrix is a table of frequencies or unconditional probabilities.

**Expected Values for Kappa Given Fallible Observers**

Now we can compute the expected unconditional probabilities for the cells of the agreement matrix given fallible observers. The formula is

\[ u_{ij} = \sum_k p_{ik} \sigma_{jk} \pi_k, \tag{1} \]

where \( u_{ij} \) represents a cell in the \( K \times K \) agreement matrix. Each \( u_{ij} \) is the sum of \( K \) terms, where each term represents the probability that the first observer will code an event \( C_i \) and the second observer will code it \( C_j \) given a \( C_k \) event. Per basic probability theory, the probability of the joint event that constitutes each term is a product (and), whereas the probability of any of these joint events occurring is a sum (or). The terms in each series exhaust the possible ways the first observer might code an event \( C_i \) when the second observer codes it \( C_j \). Consider \( u_{11} \), which is the probability that both observers will classify an event \( C_1 \). When \( K = 3 \),

\[ u_{11} = \rho_{111} \sigma_{111} \pi_1 + \rho_{112} \sigma_{112} \pi_2 + \rho_{113} \sigma_{113} \pi_3. \]

The first term indicates that both observers correctly classified a \( C_1 \) event, whereas the second and third terms indicate that what appears to be agreement can occur when both observers misclassify an event the same way.

A numeric example may clarify Equation 1 further. Let \( K = 2 \),

\[ \rho = \sigma = \begin{bmatrix} .90 & .10 \\ .15 & .85 \end{bmatrix}, \quad \text{and} \quad \pi = \begin{bmatrix} .125 \\ .875 \end{bmatrix}. \]

The brackets enclose the \( K \times K \) matrix of \( \rho \) and \( \sigma \) conditional probabilities and the \( K \) simple probabilities. In this case, the simple probabilities are highly variable (\( \pi_1 = .25/K \)) and both observers code \( C_1 \) and \( C_2 \) events correctly 90% and 85% of the time, respectively (\( \rho_{111} = \sigma_{111} = .90, \rho_{212} = \sigma_{212} = .85 \)). As a result, the probability that both observers will assign code \( C_1 \) to an event is .113 (i.e., \( u_{11} = \rho_{111} \sigma_{111} \pi_1 + \rho_{112} \sigma_{112} \pi_2 = .90 \times .90 \times .125 + .15 \times .15 \times .875 = .101 + .012 = .113 \)).

Using Equation 1, we computed expected values of kappa, systematically varying \( K \), variability of simple probabilities, and observer accuracy (kappa can also be computed directly from \( \rho, \pi \), and \( \sigma \); see the Appendix). We let levels of observer accuracy range from 80% to 100%: 80% is a value that Gardner (1995) characterized as discouragingly low, "but possibly representative of the accuracy of classification for some social behaviors or expressions of affect" (p. 347), and is one we regard as a lower bound of acceptability. We made the simplifying assumptions that for any one kappa computation all codes were detected with equal accuracy (i.e., all diagonal probabilities were the same) and inaccuracy (i.e., all off-diagonal probabilities were the same) and that both observers were identically fallible (i.e., \( \rho_{ij} = \sigma_{ij} \) for all \( i \) and \( j \)). These assumptions are unlikely to be met perfectly in practice, yet we believe computations based on them can be useful. When these assumptions are reasonable and investigators believe that the accuracy with which each code is applied meets or exceeds a given minimum level, then our computations, which assume that the accuracy for all codes only meets that minimum level, should provide a lower-bound estimate for expected values of kappa.
Results of Kappa Computations

Our results are shown in Figure 1. The figure contains four sets of lines, where successively heavier lines represent \( K = 2, 3, 5, \) and 10, respectively. Within each set, the top line represents equiprobable, the middle line moderately variable, and the bottom line represents highly variable simple probabilities (distinct lines are not visible when \( K = 10 \)). The computations depicted in Figure 1 suggest that the interpretation of kappa depends on the number of codes and, especially when codes are few, on the variability of their simple probabilities. Others have noted that, even when accuracy is quite good, kappas can be quite low when simple probabilities are grossly unequal (e.g., Grove, Andreasen, McDonald-Scott, Keller, & Shapiro, 1981; Kraemer, 1979) but have focused primarily on \( 2 \times 2 \) tables where rows and columns represent presence or absence of a diagnosis (an exception is Umesh, Peterson, & Sauber, 1989). Our computations extend earlier work to consider several values of \( K \) and degree of simple probability variability.

As can be deduced from Figure 1, when circumstances reflect the assumptions made here and \( K = 5 \) or more, (a) results are little affected by the variability of the simple probabilities and (b) interobserver kappas in the mid to high .50s, .60s, .70s, and .80s suggest observer accuracy for simple events near .80, .85, .90, and .95, respectively. However, when \( K = 2 \) and codes are equiprobable, kappas near .35, .50, .65, and .80 suggest observer accuracies near .80, .85, .90, and .95, respectively; and when the two codes are highly variable, observer reliabilities near .80, .85, .90, and .95 are suggested by kappas as low as .20, .30, .44, and .65, respectively. Thus, depending on circumstances, quite reasonable observer accuracy for simple events (e.g., \( p_{ij} = .90 \) for all \( i \)) may yield quite dissimilar, and occasionally quite low, values of kappa. As a result, interpretations of kappa—including definitions of what constitutes a good kappa—should take these circumstances into account.

As noted earlier, our computations assumed that coders were equally accurate and inaccurate for all codes. When investigators believe this assumption is untenable, and have some basis for providing values for \( \rho \) and \( \sigma \), they can use Equation 1 or the formula in the Appendix (or a computer program we have developed\(^1\)) to compute expected values of kappa tailored to their circumstances. Still, to gain some sense of how expected kappas might be affected when \( p_{ij} \) is not the same for all \( i \), we computed expected values of kappa for the usual four values of \( K \) and three levels of simple probability variability, but this time let \( p_{ij} = .70 \) for one code and .90 for the others. We let each code in turn be the less accurate code so that we could note the effect of a less accurate code being either quite rare or quite common (when simple probabilities varied).

Results suggested that Figure 1, or the formulas from which they were generated, might be used, albeit cautiously, to estimate average accuracy from observed kappas. For example, when \( K = 5 \), simple probabilities were moderately variable, and observers were 70% accurate for one code and 90% accurate for the rest, expected kappas ranged from .65 to .72 depending on how rare the less reliable code was; when \( \pi = .20 \) for the less reliable code, the expected value

---

\(^1\) We have developed a computer program that, given user-specified values for \( \rho, \sigma, \) and \( \pi \), computes expected values for kappa, and given user-specified values for \( \rho, \pi, \) and \( \tau \), computes latent and manifested expected values for Yule's Q. The program is named FallIObs, was written in Borland's Delphi 2, and requires Windows 95. It can be downloaded from [www.gsu.edu/psychology/bakeman](http://www.gsu.edu/psychology/bakeman).
was .68. Using Figure 1 and interpolating, a kappa of .68 suggests an observer accuracy of 86% (remember, \( K = 5 \) and probabilities were moderately variable), yet in this case we know that four of the accuracies are 90% and one is 70%, which indeed averages 86% exactly. This single example is not proof, of course, but it is consistent with other examples we examined. When assumptions of equal accuracy and inaccuracy are reasonably met, we think Figure 1 can provide suggestive estimates of observers' average accuracy for simple events reflected by a particular value of kappa. Investigations interested in estimating observer accuracy under these assumptions for other values of \( K \) or probabilities of simple events can use Equation 9 in the Appendix. However, when assumptions are questionable, investigators should use Equation 1 or Equation 1A in the Appendix (or our computer program) instead.

Fallible Observers and Sequential Patterns

In the previous section, we saw that expected values for kappa calculated to assess the extent to which observers agree when coding simple events varied, depending on factors such as number of codes and the variability of their simple probabilities, even when observers maintained a constant level of accuracy. In this section we consider the influence of similar factors on the ability of fallible observers to identify sequential patterns.

The classical psychometric distinction between true and error components of measured scores may not remain fresh in the minds of practicing researchers. Gardner (1995) argued that this may be especially true of those who analyze sequential categorical data. Such researchers, he suggested, often fail to grasp the magnitude of the problem that inaccurate measurement can cause. Gardner recommended that researchers explicitly think in terms of an underlying sequential process that generates latent sequential states and a measurement process that records manifest sequential states because this serves to remind us that measurement error may be present. He then demonstrated that, especially when codes are few and measurement of simple events not sufficiently accurate, estimated transitional probabilities—one of the more common sequential statistics—can give a quite misleading picture of the true latent process.

In this section, we extend Gardner's argument to a wider range of situations to determine if some circumstances are more vulnerable than others to unreliable measurement. For simplicity, we focus on transitions from one event to a second, immediately following event; thus the lag between coded events is assumed to be 1 although arguments presented here generalize to other lags. In principle, we might be concerned with any one or several of the \( K^2 \) possible two-event sequences involving the \( C_1 \) through \( C_K \) codes, but for generality and ease of exposition, we usually designate one of the codes \( A \), another \( B \) (both could be the same code), and then discuss the \( A \) to \( B \) transition.

Indices of Sequential Pattern

Counts of two-event sequences are usually presented as a \( K \times K \) table of transitional frequencies. When \( K \) is not 2 initially, often we collapse such matrices into \( 2 \times 2 \) tables, where the first row represents the antecedent event \( A \), the second row all antecedent events that are not \( A \), the first column the consequent event \( B \), and the second column all consequent events that are not \( B \); thus the upper left-hand cell contains the observed frequency for \( A \) to \( B \) transitions. A statistic on which we often rely is Yule's Q, an index for \( 2 \times 2 \) tables that is derived from the odds ratio but algebraically transformed so that, like the familiar Pearson product-moment correlation, it varies from \(-1\) to \(1\) (see Bakeman, McArthur, & Quera, 1996). Negative values indicate a transitional probability that is less than expected and positive values indicate a transitional probability that is greater than expected, given the base rates for its constituent codes; thus a large positive value suggests that the \( A \) to \( B \) transition is a recurring pattern, characteristic of the data analyzed.

Other statistics, such as the phi coefficient, which is monotonically invariant with Yule's Q, could also be used and would give essentially similar results, but Yule's Q is somewhat less restricted and so was used here (see Bakeman, McArthur, & Quera, 1996). We do not consider a \( z \) score (for \( 2 \times 2 \) tables, \( z = \sqrt{n_{AB}n_{BA}} \); see Bakeman & Quera, 1995; for a slightly different formula, see Allison & Liker, 1982) because, unlike Yule's Q, its value is affected by the number of tallies as well as the strength of the transition. We consider the transitional probability (e.g., the probability that the second event will be coded \( B \) given that the first event was coded \( A \), symbolized \( p_{BA} \)) only secondarily; although informative descriptively, its value does not reveal how strongly the observed probability of an \( A \) to \( B \) sequence deviates from its expected value, on the basis of simple probabilities.
**Gardner's Formulas**

The foundation for Gardner’s (1995) argument is as follows. Assuming a set of $K$ codes designated $C_1$ through $C_K$, let $\pi_i$ represent the true probability for a $C_i$ event. Additionally, let $\tau_{ij}$ represent the true transitional or conditional probability (i.e., the probability that a $C_j$ event will occur given a preceding $C_i$ event), $\gamma_{ij}$ the true unconditional probability (i.e., the probability of a $C_i$ to $C_j$ sequence), $g_{ij}$ the unconditional probability for a manifest sequence (i.e., the probability of coding a $C_i$ to $C_j$ sequence), and $\rho_{ij}$ the reliability of event measurement, that is, observer accuracy as discussed in the previous section.

Given perfect measurement, the unconditional probability that a randomly sampled transition will be coded $C_i$ to $C_j$ is

$$g_{ij} = \gamma_{ij} = \pi_i \tau_{ij}.$$  

(2)

Given fallible measurement, however, the unconditional probability that a randomly sampled transition will be coded a $C_i$ to $C_j$ sequence is

$$g_{ij} = \sum_{r=1}^{K} \sum_{s=1}^{K} \rho_{ir} \pi_i \tau_{ir} \rho_{sj}.$$  

(3)

Equation 3 represents expected values for the unconditional probability of coding a $C_i$ to $C_j$ sequence when codes are measured unreliable. Thus expected values for the observed transitional probabilities are

$$t_{ij} = \frac{g_{ij}}{\sum_j g_{ij}}.$$  

(4)

These formulations derive from Gardner (1995, pp. 344–345) but are expressed in notation that may be more familiar and accessible to many psychologists.

An example can clarify Equation 3 and, at the same time, demonstrate the problems unreliable measurement can create. Assume, for example, that $K = 2$,

$$\rho = \begin{bmatrix} .8 & .2 \\ .2 & .8 \end{bmatrix}, \pi = \begin{bmatrix} .125 \\ .875 \end{bmatrix}, \text{ and } \tau = \begin{bmatrix} .560 & .440 \\ .063 & .937 \end{bmatrix},$$

where brackets enclose the $K \times K$ matrices of $\rho$ and $\tau$ conditional probabilities and the $K$ simple probabilities. For this example, $C_1$ and $C_2$ events are both coded with 80% accuracy ($\rho_{11} = \rho_{22} = .8$), the simple probabilities are highly variable ($\pi_1 = .125$, $\pi_2 = .875$), and $C_1$ is reciprocated somewhat more often than not ($\tau_{11} = .560$), whereas $C_2$ is almost always reciprocated ($\tau_{22} = .937$).

These particular transitional probabilities were selected so that the latent value for Yule’s $Q$, which is computed as

$$\frac{\gamma_{11} \gamma_{22} - \gamma_{12} \gamma_{21}}{\gamma_{11} \gamma_{22} + \gamma_{12} \gamma_{21}}$$  

(5)

would equal .90, thus representing an unusually large effect (for derivation of the formula for Yule’s $Q$, see Bakeman, McArthur, & Quera, 1996). For this example, the true unconditional probabilities per Equation 1 are

$$\gamma = \begin{bmatrix} .070 & .055 \\ .055 & .820 \end{bmatrix}$$

whereas the expected values for the observed unconditional probabilities per Equation 3, given fallible observers, are

$$g = \begin{bmatrix} .095 & .180 \\ .180 & .545 \end{bmatrix}.$$

Thus, expected values for the observed transitional probabilities per Equation 4 are

$$t = \begin{bmatrix} .346 & .654 \\ .248 & .752 \end{bmatrix}$$

(i.e., .095/.275, .180/.275, .180/.725, and .545/.725). This example is similar to one presented by Gardner (1995, pp. 346–347).

The computational matrix for $g$ given in Figure 2 illustrates how Equation 3 works. Insofar as reliabilities for correctly identifying simple events are high, then the expected values for $g$ will mirror values for $\gamma$ (and hence expected values of $t$ will mirror those for $\gamma$) and the bolded lines in each cell of the computational matrix shown in Figure 2 will largely determine the expected value of $g$ for that cell. As an example,

| \begin{bmatrix} .8 & .125 \\ .8 & .125 \end{bmatrix} \times \begin{bmatrix} .560 & .440 \\ .063 & .937 \end{bmatrix} \times \begin{bmatrix} .8 \end{bmatrix} | = \begin{bmatrix} .8 \times .125 \times .440 & .009 \\ .8 \times .125 \times .063 & .009 \end{bmatrix},  
| \begin{bmatrix} .8 & .125 \\ .8 & .125 \end{bmatrix} \times \begin{bmatrix} .560 & .440 \\ .063 & .937 \end{bmatrix} \times \begin{bmatrix} .8 \end{bmatrix} | = \begin{bmatrix} .8 \times .125 \times .440 & .009 \\ .8 \times .125 \times .063 & .009 \end{bmatrix}.

Figure 2. Computational matrix for observed unconditional probabilities per Equation 3 assuming the values for $\rho$, $\pi$, and $\tau$ given in the text. Computations were performed before rounding so displayed numbers may not sum exactly.


consider the value of the expected unconditional probability for coding a C₁ to C₂ transition (c₁₂ = .180). The four lines in the upper right-hand cell of Figure 2 represent four ways transitions might be coded C₁ to C₂: (a) the first event is correctly coded and the second is incorrectly coded (i.e., the second event is not truly C₂), (b) both are correctly coded, (c) both are incorrectly coded, and (d) the first is incorrectly coded and the second is correctly coded. Per basic probability theory, the probability for each possibility is the product of its constituent parts. For example, the probability for the first possibility (C₁ is correct but C₂ is incorrect) is the probability of correctly identifying the first event as C₁ (p₁₁ = .8) times the latent probability of a C₁ event (π₁₁ = .125) times the latent probability of the second event being a C₁ given that the first event was a C₁ (τ₁₁ = .560) times the probability of incorrectly identifying the second event as C₂ when it is actually C₁ (p₂₁ = 0.2). In this case, the probability for this possibility is relatively low (.011), but the probability for the fourth possibility (C₁ is incorrect but C₂ is correct) is relatively high (.131), almost four times more likely than the desired second possibility (.035, both correctly coded).

The expected values for r in this case are quite disconcerting. In the latent process, given an antecedent C₁, the probability of a C₁ following was greater than the probability of a C₂ following (i.e., τ₁₁ > τ₂₁, .560 > .440). Yet for the manifest process, given an antecedent C₁, the probability of a C₁ following was less than the probability of a C₂ following (i.e., τ₁₁ < τ₂₁, .346 < .654). As Gardner (1995) noted for his similar example, when latent simple probabilities are highly unequal (here π₁ = .125, π₂ = .875) and effects strong (here Yule’s Q = .90), latent-to-manifest reversals can occur among expected transitional probabilities (here τ₁₁ > τ₂₁ but f₁₁ < f₂₁).

Other statistics were degraded but not reversed. For example, whereas

\[ \pi = \begin{bmatrix} .125 \\ .875 \end{bmatrix}, p = \begin{bmatrix} .275 \\ .725 \end{bmatrix} \]

in other words, given fallible observers the expected observed (manifest) simple probabilities became less unequal than the true (latent) ones. Similarly, whereas Yule’s Q for the latent process was a strong .90, the value for the observed process was .23, still positive but considerably diminished. In general, as the accuracy with which events are coded (i.e., the pᵢᵢ elements on the diagonal of the p matrix) decreases from 1 to chance, manifest simple probabilities tend to become equiprobable (i.e., 1/K) and values for Yule’s Q tend to zero. In one sense, this is encouraging: As observers’ judgments become more random, no association may be detected where one exists (i.e., values of Yule’s Q → 0) but at least the nature of the association does not reverse (i.e., a negative Yule’s Q is not observed when the latent Yule’s Q is positive).

**Results of Yule’s Q Computations**

A fall in the index of association from .90 to .23 seems quite serious. To determine whether such deterioration is rare or routine, we systematically varied several factors. We assumed an interest in two-event sequences and, for generality, referred to the first code in the sequence as A and the second as B, where A represents one of the codes C₁ through C₅; likewise for B. Then we computed values for the Yule’s Q associated with an A to B transition, first based on perfectly measured and then on varying degrees of imperfectly measured data. Four factors were varied: the number of codes (i.e., K), the variability of their simple probabilities, observer accuracy, and the size of the transitional effect.

As in the previous section, we let K = 2, 3, 5, and 10 and, for each value of K, defined three levels of simple probability variability. We made the simplifying assumption that for any one Yule’s Q computation all codes were detected with equal accuracy (i.e., all p diagonal probabilities were the same) and inaccuracy (i.e., all p off-diagonal probabilities were the same) and let levels of observer accuracy range from 80% to 100%. Additionally, we defined small, medium, and large transitional effects as those resulting in values of Yule’s Q of .25, .50, and .75, where the size of the effect is relative to the A to B transition. For simplicity and ease of exposition, we focus on just one of the K² possible transitions (the A to B transition), although in practice any one or several of the transitions may be of interest to investigators.

What remains is specification of the codes to represent the general codes A and B. We chose C₁ for both, where C₁ is the least frequent code when simple probabilities are either moderately or highly variable. Given our assumptions regarding p and granted a given level of observer accuracy, expected observed values for individual transitional probabilities (Equation 4) vary as a function of the simple probabilities and the strength of the effect. Thus when simple probabilities are equiprobable, it does not matter which codes we select for A and B. When simple probabilities varied, we chose the least frequent code to rep-
resent A and B because doing so showed greater deterioration in Yule's Q for a given set of circumstances than selecting more frequent codes (this can be demonstrated using the computer program mentioned in Footnote 1). Our intent in performing these computations was to determine whether deterioration of the sort Gardner (1995) demonstrated was routine over a variety of circumstances, and for this purpose selecting a worst-case scenario establishes a baseline. When other assumptions are met but investigators are concerned with transitions between more frequent codes, then the computations reported here for the A to B transition should provide a conservative estimate for expected deterioration in Yule's Q, and when our assumptions seem unreasonable, investigators can use Equations 3-5 (or the computer program mentioned in Footnote 1) to compute true and observed values of Yule's Q tailored to their circumstances.

Results for latent values of Yule's Q = .75, .50, and .25 are displayed in Figures 3-5, respectively. Each figure contains four sets of lines, where successively heavier lines represent K = 2, 3, 5, and 10, respectively. Within each set, the top line represents equiprobable and the bottom line highly variable simple probabilities; for simplicity, no line for moderately variable probabilities is shown but it would have fallen in between. Deterioration in the value of Yule's Q was particularly pronounced for the circumstances that Gardner (1995) investigated—K = 2, highly variable simple probabilities, and low observer accuracy—but was less as K increased, codes were more nearly equiprobable, and accuracy levels increased. Additionally, deterioration appeared greater for smaller manifest values of Yule's Q (compare Figures 3-5).

Still, for the circumstances indicated in Figures 3-5
reversals of the sort that troubled Gardner (1995; e.g., $\tau_{111} > \tau_{211}$ yet $f_{111} < f_{211}$) occurred only when $K = 3$, $\pi_1 = .167$, and accuracy $\leq .875$ and when $K = 2$, $\pi_1 = .25$, and accuracy $\leq .95$. Such reversals became more likely when the effect was strong (e.g., $\tau_{111}$ was much greater than $\tau_{11}$) especially when $K$ was small. For example, for Yule’s $Q = .90$, reversals occurred when $K = 10$, $\pi_1 = .025$, and accuracy $\leq .825$; when $K = 5$, $\pi_1 = .05$, and accuracy $\leq .90$; when $K = 3$, $\pi_1 = .083$, and accuracy $\leq .95$; and when $K = 2$, $\pi_1 = .125$, and accuracy $\leq .975$ and $\pi_1 = .25$, and accuracy $\leq .85$.

Investigators should always strive for highly accurate observers, of course, but accuracy becomes especially important when codes are few (e.g., $K = 2$ or 3), simple probabilities are highly variable, effects of interest are strong, and patterns involving the least frequent codes are of interest. Then observer accuracy for single events as high as .90 or .95 may be required. However, as our computations show, under a fairly wide range of less extreme circumstances the ability to detect patterns of interest is compromised but not crippled by fallible observers. When codes are many and roughly equiprobable, observer accuracy for single events as low as 80% may be acceptable under circumstances specified later.

Reliability for Sequential Patterns

Kappa has become the statistic of choice for demonstrating interobserver agreement and is often presented as evidence of observer reliability. Conventionally, kappa is used to gauge agreement with respect to simple events although Gardner (1995) suggested that kappa might also be used to gauge agreement with respect to patterns of events. Per his suggestion, rows and columns of the agreement matrix would be labeled, not with simple codes, but with two-code sequences. For example, when $K = 2$, rows and columns of the agreement matrix would be labeled $C_1 C_1$, $C_1 C_2$, $C_2 C_1$, and $C_2 C_2$. However, the resulting agreement matrices would consist of $K^2$ instead of $K^4$ cells, which rapidly becomes unwieldy as $K$ increases. This circumstance led us to consider a more traditional and simpler approach. If investigators’ substantive concerns rest with particular transitions rather than single events, a statistic such as Yule’s $Q$ may be used to index the extent to which particular transitions are characteristic of a given data set. Thus it makes sense to determine the reliability of these indices, as determined by the usual formal reliability study (e.g., Fleiss, 1986; Wiggins, 1973).

Examining both kappa and formal reliability statistics constitutes a two-step approach with considerable merit. First, investigators compute kappa for simple codes because the value of kappa and the agreement matrix on which it is based provide useful feedback for training observers (Bakeman & Gottman, 1986, 1997) and because accurate coding of simple events is necessary for the accurate coding of event sequences. Second, investigators compute the reliability of a sequential statistic such as Yule’s $Q$ because the sequential statistics are the ones analyzed when attempting to answer substantive sequential questions. In other words, sequential indices can be treated as any other score, and traditional psychometric methods can be used for estimating their reliability. This has the merit of establishing reliability for the scores ultimately analyzed (Bakeman, in press).

Traditionally, some form of intraclass correlation or generalizability coefficient is used to estimate reliability (Wiggins, 1973). Let $S$ indicate the number of sessions coded for the reliability study and $O$ the number of observers who coded them (observers code each session independently). Then, if a Yule’s $Q$, for example, is computed for the $A$ to $B$ transition for $S$ sessions by $O$ observers, the result is an $S \times O$ matrix of scores, arranged as for a repeated measures analysis. Given such scores, an appropriate coefficient of reliability is

$$\alpha_o = \frac{MS_S - MS_S \times O}{MS_S + (O - 1) \times MS_S \times O}, \quad (6)$$

where $MS_S$ and $MS_S \times O$ are the mean squares for sessions and the $S \times O$ interaction (i.e., the error term for a repeated measures analysis of variance, Wiggins, 1973). 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 correlation is defined as

$$\alpha_o = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2}. \quad (7)$$

(Equation 6 is derived from 7 by substitution and algebraic manipulation.) This statistic, symbolized $\alpha_o$ here to avoid confusion with the more common index of internal consistency usually called Cronbach’s alpha, estimates the reliability of observations made by a randomly selected observer, selected from the pool that contained the two observers used for the reliability study, and further assumes that data will be inter-
interpreted within what Suen (1988) termed a norm-referenced (i.e., values are meaningful only relatively as with rank-order statistics like correlation coefficients) as opposed to a criterion-referenced framework (i.e., interpretation of values references an absolute external standard as with statistics like unstandardized regression coefficients). Equation 6 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, although, as a practical matter, values for various intraclass correlation coefficients may not differ greatly.

**Simulating a Reliability Study**

To understand the effects of fallible observers on $\alpha_\omega$, we performed a computer simulation. As earlier, we assumed two observers who detected all codes with equal accuracy and inaccuracy, and systematically varied values for the number of codes, the variability of their simple probabilities, and the accuracy with which simple events were coded. For each combination of factors, we simulated 10 reliability sessions. For each session, a random latent sequence and two random manifest sequences (one for each observer) were generated, all constrained by parameter values for $K$, $\pi$, $p$, and $\sigma$, as detailed shortly. The size of the transitional effect, which we assumed was normally distributed, varied randomly for each session.

The remaining variable was sequence length. Sequence length is important because, other things being equal, the reliability of sequential statistics increases with length. Indeed, the minimum length that sequences might safely assume is an important matter for investigators to consider. As a general rule, shorter sequences are permissible when codes are few (i.e., $K$ is small) and their simple probabilities are roughly equal, but longer sequences are required as $K$ increases and as simple probabilities become more variable because otherwise too many expected frequencies in the lagged $K \times K$ tables may be zero or close to it. Accordingly, the length of sequences generated by the simulation program was determined by $K$ and simple probability variability. Two sets of lengths were investigated, each defined by a minimum acceptable value for the smallest expected frequency in the $2 \times 2$ table used to compute Yule's $Q$ for the $A$ to $B$ transition. This minimum value was 10 for the first set (a conservative minimum sometimes recommended for a chi-square test with 1 degree of freedom; see Hays, 1963); the lengths for various values of $K$ and variability of simple probabilities, both when the sequence under consideration was composed of the least frequent and most frequent codes, are shown in Table 2. The minimum value for the second set was 20, which resulted in lengths double those shown in Table 2.

For each simulated reliability session, a latent sequence was generated constrained by the current value for $K$, variability of simple probabilities, and a randomly determined effect. Then, for accuracy levels ranging from 80% to 100% in steps of 2.5%, manifest sequences were generated for each observer. For each pair of manifest sequences, a Yule's $Q$ for the $A$ to $B$ transition was computed (where $A = B = C_1$, the least frequent code) and then $\alpha_\omega$ was computed for the 10 reliability sessions per Equation 6. The reliability study was replicated 1,000 times and mean $\alpha_\omega$ values were computed for each combination of parameter values. Alphas were also computed for $A$ to $B$ transitions when $A = B = C_K$ (the most frequent code) but

<table>
<thead>
<tr>
<th>$K$</th>
<th>Highly variable, $\pi_A = \pi_B = .25/K$</th>
<th>Moderately variable, $\pi_A = \pi_B = .5/K$</th>
<th>Equiprobable, $\pi_A = \pi_B = 1/K$</th>
<th>Moderately variable, $\pi_A = \pi_B = 1.5/K$</th>
<th>Highly variable, $\pi_A = \pi_B = 1.75/K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>640</td>
<td>160</td>
<td>40</td>
<td>160</td>
<td>640</td>
</tr>
<tr>
<td>3</td>
<td>1,440</td>
<td>360</td>
<td>90</td>
<td>40</td>
<td>58</td>
</tr>
<tr>
<td>5</td>
<td>4,000</td>
<td>1,000</td>
<td>250</td>
<td>111</td>
<td>82</td>
</tr>
<tr>
<td>10</td>
<td>16,000</td>
<td>4,000</td>
<td>1,000</td>
<td>444</td>
<td>327</td>
</tr>
</tbody>
</table>

Note. The first two columns give sequence lengths when $\pi_A = \pi_B = \pi_1$, the least frequent code; the middle column gives sequence lengths when $\pi_A = \pi_2 = \pi_1$, but all codes are equiprobable; the last two columns give sequence lengths when $\pi_A = \pi_2 = \pi_K$, the most frequent code (see Table 1). Length is computed by dividing the minimum cell size desired (here 10) for the $2 \times 2$ Yule's $Q$ table (rows labeled $A$ and not $A$, columns labeled $B$ and not $B$) by the expected (unconditional) probability for an $A$ to $B$ sequence (or by the expected probability for a not-$A$ to not-$B$ sequence if it is smaller as occurs when $K = 2$ and codes are moderately or highly variable and when $K = 3$ and codes are highly variable). For example, when $K = 3$ and $\pi_i = .167$ (least frequent, moderately variable), length = $10/1.167^{12} = 360$. A minimum cell size of 20 would require lengths twice those shown.
are not reported here because values fell between the equiprobable and moderately varied case when \( A = B = C_1 \). Sets of 10 reliability sessions were run for all combinations of values defined by the two minimum expected frequencies (10, 20), four values of \( K \) (2, 3, 5, 10), and three values of simple probability variability (equiprobable, moderately variable, and highly variable) with one exception: When \( K = 10 \), the sequence length required when \( A = B = C_1 \) and simple probabilities are highly varied is extremely long (see Table 2).

**Results of the Simulations**

The results of the simulation are portrayed in Figures 6 and 7 for minimum expected frequencies = 10 and 20, respectively. These figures contain four sets of lines, where successively heavier lines represent \( K = 2, 3, 5, \) and 10, respectively. Within each set, the top line represents values for alpha associated with the

\[ A \rightarrow B \] transition when simple probabilities are equal (i.e., \( \pi_A = \pi_B = 1/K \)). The bottom line when \( K = 10 \) and the middle line for other values of \( K \) represent values for alpha when simple probabilities were moderately varied (i.e., \( \pi_A = \pi_B = .5/K \)) and the bottom line when \( K = 2, 3, \) and 5 represents values for alpha when simple probabilities were highly varied (i.e., \( \pi_A = \pi_B = .25/K \)).

Figures 6 and 7 suggest the levels of reliability for sequential patterns that investigators can expect, given various values of \( K \) and observer accuracy. Congruent with our results in the previous section showing how values of Yule's Q were compromised by fallible observers, effects were most severe when codes were few and patterns involved relatively infrequent codes. Under those circumstances, higher observer accuracy is required. For example, when \( K = 2 \) and minimum expected frequency = 10, observer accuracy for simple events of approximately .94-96, was required for an alpha of .70 (Figure 6); this range became .91-93 when minimum expected frequency was doubled (Figure 7). However, when codes were several, requirements for observer accuracy (as always, with respect to simple events) may be relaxed somewhat. For example, when \( K = 5 \) and minimum expected frequency = 10, observer accuracy for simple events of approximately .88-.93 was required for an alpha of .70 (Figure 6); this range became .82-.88 when minimum expected frequency was doubled (Figure 7). For the range of parameter values we investigated, observer accuracies in the low .80s
resulted in alphas of .70 or greater only when $K = 10$ and minimum expected frequency = 20, but then very long sequences may be required (see Table 2), especially when patterns of interest include relatively infrequent events.

Summary and Recommendations

Gardner's (1995) suggestion that the codes observers assign to sequential events be regarded as the manifest results of fallible measurement is extremely useful. It serves to remind us, yet again, of the central psychometric insight that measurement only provides an imperfect reflection of the underlying, latent state of affairs. Thus, as observers become less accurate—that is, as the probability that they will assign correct codes to events decreases—our view of the latent world increasingly slips out of focus.

How severe the effects of fallible observers are seems to depend, in part, on the number of codes and the variability of their simple probabilities. On the basis of the range of circumstances we considered and the assumptions we made, almost always deterioration was most severe when codes were few (e.g., two or three) and the variability of their simple probabilities quite high. Yet severity of deterioration may also depend on the sequential statistic examined. As discussed earlier, fallible observers can be more problematic when considering an observed transitional probability, which describes an effect, than for Yule’s Q, which describes the magnitude of an effect. A worst case example, presented earlier, examined reciprocity for $C_1$. When $C_1$ was relatively rare ($\frac{r_{11}}{s_{11}} = .125, K = 2$) and observers were only 80% accurate, the latent conditional probability of a $C_1$ given a previous $C_1$ was .560, suggesting slight reciprocity (i.e., $\tau_{11} > \tau_{21}$). Yet the manifest value, given fallible observers, was .346 (i.e., $t_{11} < t_{21}$), suggesting a lack of reciprocity (but see next paragraph). Such reversals rightly concerned Gardner (1995), yet our computations suggest that they occur primarily in fairly limited circumstances: When codes are few, patterns of interest involve relatively infrequent events, effects are strong, and observers are weak (e.g., less than 80% accurate). In any event, using the computations presented here investigators can determine whether or not such reversals are likely to be problematic given their specific circumstances.

In general, it makes more sense to focus on the value of Yule’s Q (or of other indices of association such as phi) associated with a particular transition than on the value of the transitional probability itself because such indices indicate whether or not the observed degree of association is greater than expected, given base rates. For the example given earlier, the expected value for $t_{11}$, based on observed base rates, was .275 (i.e., expected for $t_{11} = p_{11} = .275$). The observed value for $t_{11}$, .346, exceeded this, suggesting reciprocity. True, as noted in the previous paragraph, $t_{11} < t_{21}$, but this, although descriptively informative, reflects, in part, how seldom $C_1$ events were observed ($p_1 = .275$ vs. $p_2 = .725$).

The aspect of usual interest to investigators—for example, whether $t_{11}$ is greater than expected given base rates—is reflected by Yule’s Q, not the transitional probability, and although the observed value of Yule’s Q is degraded by fallible observers—from .90 to .23 for this worst case scenario—it is not reversed. Given a true positive association, and assuming that observers do no worse than chance, no association or a weaker positive association may be observed, but not a negative one. Often magnitude of effect statistics such as Yule’s Q are used as scores in subsequent analyses (t tests, analyses of variance, multiple regressions, etc.; see Bakeman, McArthur, & Quera, 1996; Bakeman, Robinson, & Quera, 1996). Such analyses may fail to find effects present in the latent process due to fallible observers, but they will not find effects counter to those in the latent process.

The effects of fallible observers on indices of association such as Yule’s Q are seldom discussed. In contrast, the effects of fallible observers on kappa are widely known: Fallible observers decrease kappa. Less widely appreciated is the effect of variability of simple probabilities on values of kappa. What is usually termed the base rate problem (Grove et al., 1981; Kraemer, 1979)—precipitous declines in kappa when agreement is less than perfect and simple probabilities are quite variable—has been addressed largely in the context of 2 x 2 tables reflecting presence or absence of a diagnostic category. Here we extend discussion to situations involving more than two codes and different degrees of variability in probabilities of simple codes.

Our computations suggest no one value of kappa can be accepted as adequate, as convenient as this might be. Instead, adequate values depend on circumstances (e.g., the number of codes and the variability of their simple probabilities). Still, our computations do suggest a principled way for identifying adequate values of kappa for various circumstances. For example, although shortly we suggest a rationale for selecting a desired level of accuracy, assume for now...
that we have defined an adequate kappa as one that results from observers who are 90% accurate. Then, assuming that observers are equally accurate and inaccurate over all codes, when \( K \) is 2 kappas between .45 and .65 might be acceptable depending on the degree of simple probability variability, when \( K \) is 3 values falling between approximately .68 and .73 might be acceptable, and when \( K \) is 5 or greater kappas should exceed .75 (see Figure 1). Guidance of this sort is often desired by investigators and, to our knowledge, is not provided elsewhere; here it is offered with the caveat that the guidelines apply only to the extent that assumptions regarding \( p \) and \( \sigma \) are met.

The simulations we performed provide further guidance for investigators. First, to perform the simulations at all, we had to select an appropriate sequence length. The rationale we developed was based on a minimum expected cell frequency for summary tables and was affected by the number of codes and the variability of their simple probabilities (see Table 2). In a manner analogous to power analysis, this rationale can also be used by investigators when first determining how much data they should collect (and some investigators may believe that the value of 10 we used is too conservative; e.g., see Wickens, 1989). Quite long sequences, even in excess of a thousand events, may be required when codes are many and patterns of interest involve relatively infrequent codes. Additionally, more data are better, in the sense that all other things being equal longer sequences result in greater reliability for sequential statistics (compare Figures 6 and 7).

Second, the simulations highlight a relatively unused but appealing approach to reliability in sequential studies, not at the level of the data collected but at the level of the data analyzed. From this point of view, kappa is useful primarily when training observers and monitoring data collection. However, once data are collected and analysis is about to begin, then it makes sense to establish the reliability of the sequential indices actually analyzed, such as Yule’s Q. True, this could be seen as problematic in an exploratory study. For example, if \( K \) were 20 then 400 possible Lag 1 patterns exist, and establishing the reliability of each pattern seems excessive. But requiring investigators to define a handful of patterns of particular interest on which analytic and interpretative energy focuses on particular patterns. These levels, in turn, can be used to define minimum acceptable levels of kappa for a particular investigation. For example, assume that an alpha of .70 is regarded as acceptable, that \( K = 5 \), and that codes are roughly equiprobable. Then, assuming a minimum expected frequency of 10 (and so a sequence length of 250; see Table 2), observer accuracy should be no less than .88 (see Figure 6). Assuming that both observers are equally accurate and inaccurate for all events, an accuracy of .88 results in an expected kappa of approximately .73 (see Figure 1). Under these circumstances (i.e., \( K = 5 \), equiprobable codes, desired reliability for Yule’s Q at least .70), investigators should regard values of kappa less than .73 as unacceptable.

One final comment, the present discussion has considered the effects of fallible observers when investigators search for patterns in sequential data and has suggested how minimum standards for kappa might be set. Once \( K \) is set and investigators have some sense of the variability among simple probabilities, if they then select an appropriate sequence length and establish a minimum acceptable level of reliability for sequential patterns, the results of the simulations performed here can be used to suggest what level of observer accuracy is required. These results, in turn, can be used to determine minimum acceptable levels for kappa. Such guidance permits minimum levels of kappa to be set in a principled way, and seems an advance over current practice. Our guidelines, as noted earlier, depend on a number of assumptions, especially with regard to observer accuracy. Where these seem unreasonable, investigators could run similar simulations using different parameters tailored to their circumstances.

The simulation program was written in Borland’s Pascal 6.0 and runs under DOS or Windows. Interested readers may request a program listing; modifying and running the program requires some knowledge of Pascal.

References


Bakeman, R. (in press). Behavioral observations and coding. In H. T. Reis & C. K. Judd (Eds.), Handbook of
Appendix

Computing Kappa and Accuracy

Computing $\kappa$ From $\pi$, $\rho$, and $\sigma$

Kappa can be computed directly from $\pi$, $\rho$, and $\sigma$. After substituting the expected unconditional probabilities for the cells of the agreement matrix per Equation 1 into the standard formula for unweighted kappa (Cohen, 1960) and simplifying the algebra, the equation is

$$
\kappa = \frac{\sum_{i} \sum_{k} \rho_{ik} \sigma_{ik} \pi_{i} - \sum_{i} [\sum_{k} \rho_{ik} \pi_{i} - \sum_{k} \sigma_{ik} \pi_{i}]}{1 - \sum_{i} [\sum_{k} \rho_{ik} \pi_{i} - \sum_{k} \sigma_{ik} \pi_{i}]}.
$$

(1A)

This equation is general and does not require that observers be equally accurate and inaccurate for all codes or that observers perform the same.

Computing Observer Accuracy From $\kappa$ and $\pi$

Granted the assumption that an observer is equally accurate and inaccurate for all codes, accuracy ($\rho_{ii} = a$ for all $i$) can be expressed as

$$
a = \frac{s - \sqrt{s^2 - sK[\kappa(K-1)(K-2) + (1-w)(K-1)}}{sK}.
$$

(2A)

where

$$
w = \sum_{i} \pi_{i}^2 \text{ and } s = \kappa(1-wK) - K(1-w).
$$

Received June 12, 1996

Revision received March 5, 1997

Accepted March 19, 1997