Behavior Research Methods, Instruments, & Computers
1996, 28 (3), 446–457

Detecting group differences in sequential association using sampled permutations: Log odds, kappa, and phi compared

ROGER BAKEMAN and DUNCAN McARTHUR
Georgia State University, Atlanta, Georgia

and

VICENÇ QUERA
University of Barcelona, Barcelona, Spain

When determining whether a particular transition is more characteristic of one group than of another, two things are required: an index associated with the transition of interest and a statistical test that can determine whether group membership systematically affects values for that index. Here the familiar parametric *t* test is compared with a test based on sampled permutations. Indices considered are the odds and log odds ratio, Yule's *Q*, Wampold's (1989) transformed kappa, and phi. The odds and log odds ratio are monotonically increasing functions of Yule's *Q* and so give similar results. Yule's *Q* and phi are essentially rank order invariant and usually give similar results. Transformed kappa, however, rank orders subjects somewhat differently than the others; moreover, it appears somewhat biased. With respect to the tests, when subjects are 20 or more it does not matter much whether sampled permutation or parametric *t* tests are used; both yield essentially the same result. However, when subjects are fewer than 20, or whenever there is any other reason to think that parametric assumptions may not be met, permutation tests are recommended. A computer program that effects such tests is described.

In this article we review a number of relatively standard indices of sequential association and consider whether some might be better than others when testing for group differences using permutation tests. Investigators who study social interaction often define events of interest (e.g., partner complaining, daughter agreeing, toddler engaging in parallel play, infant gazing, mother responding, etc.) and ask coders to identify them in the stream of behavior (Bakeman & Gottman, 1986; Bakeman & Quera, 1992). Substantive questions usually involve sequences (or sometimes co-occurrences) of the behavioral events identified. Commonly, two-dimensional contingency tables are used to organize such data (Bakeman & Quera, 1995a; Castellan, 1979). Rows represent given events, columns represent target events, and occurrences of the joint events defined by them are tallied. For example, rows might represent the function of one person's behavior and columns the function of the partner's next (lag 1) behavior. Then transitions from the first to the second person would be tallied, and the cells would contain counts for each of the possible transitions (given questions of cross influence, autocorrelation should also be considered, as discussed later).

For present purposes, we assume that one transition is of primary theoretical interest (although comments made here generalize to other transitions in the table as well), that the study includes several subjects or experimental units (which might be individual persons, dyads, family, or other groups, etc.) assigned to two groups (although comments made here generalize to more complex designs), and that the primary question concerns whether the two groups differ with respect to the transition of interest. For example, an investigator might want to determine whether complain-complain chains are more characteristic of clinic than of non-clinic couples, or whether mothers who have received a particular intervention are more likely to respond to their infants' behavior.

In order to answer questions like these, two things are required: an index associated with the transition of interest and a statistical test that can determine whether group membership systematically affects values for that index. Some indices can be ruled inappropriate at the outset. For example, to determine whether tallies for a particular transition exceed the expected, some sort of *z* score is often proposed (Allison & Liker, 1982; Bakeman & Quera, 1995b; Fagen & Mankovich, 1980). However, contrary to advice offered earlier (Bakeman & Gottman, 1986), such *z* scores should not be used when testing for group dif-

We would like to thank Paul Yoder and Jon Tapp for a stimulating discussion that led to this article, and Byron Robinson, Richard Lambert, and Donna Borkman Reed for their helpful comments and suggestions. Special thanks are due Kevin Baldwin, who provided insightful comments on an earlier draft. Correspondence concerning this article should be addressed to R. Bakeman, Department of Psychology, Georgia State University, Atlanta, GA 30303 (e-mail: bakeman@gsu.edu) or V. Quera, Departamento de Metodologia, Facultad de Psicologia, Universidad de Barcelona, Campus Valle de Hebron, Paseo Valle de Hebron 171, 08035 Barcelona, Spain.

Copyright 1996 Psychonomic Society, Inc.

ferences. The z score is affected by the number of tallies (if the number of tallies doubled but the association remained the same, the z score would increase), and so is not comparable across experimental units (subjects, dyads, families, etc.) unless the total number of tallies remains the same for each. Some measure that is unaffected by the number of tallies, such as a strength of association or effect size measure, should be used instead (Wampold, 1992).

Strength of association or effect size measures are especially well developed for 2×2 tables (to give just two examples from an extensive literature, see Conger & Ward, 1984, and Reynolds, 1984). This is fortunate, because when interest centers on one cell in a larger two-dimensional table, the larger table can be collapsed into a 2×2, and statistics developed for 2×2 tables can be used (as Morley, 1987, noted with respect to phi). Assume, for example, that we want to know whether event B is particularly likely after event A. In this case, we would label rows A and $\sim A$ and columns B and $\sim B$ (where rows represent lag 0, columns lag 1, and ~ represents *not*). Then the collapsed 2×2 table can be represented as

	B	$\sim B$
A	a	b
$\sim A$	c	d

where individual cells labeled a, b, c, and d, as shown, represent cell frequencies.

Here we consider several indices appropriate for 2×2 tables like these and ask whether some might serve better than others when testing for group differences using sampled permutations. We prefer to base statistical significance on permutation tests rather than the more familiar parametric t test (or other parametric tests) because, first, permutation tests do not require parametric assumptions (so questions concerning their tenability do not arise) and, second, comparisons among indices are simplified when permutation instead of parametric tests are at issue, as discussed shortly.

Specifically, we compare the odds ratio, the log odds ratio, Yule's Q, Wampold's (1989) transformed kappa, and the phi coefficient. At first we regarded the parametric t test as the most obvious way to test for group differences, and so thought some indices might better meet the required assumptions than others, or that attributes such as efficiency (small standard errors of sampling distributions) and consistency (smaller standard errors as sample size increases)—attributes of estimators that matter for parametric tests (e.g., see Hays, 1963)—might distinguish among these statistics and so provide a reason for preferring one over another. However, when analyzing for group differences with permutation tests, comparison is simplified; indices that rank order subjects in the same way produce identical results, thus if rank-order invariance across subjects is established for any two indices, any additional differences need not be of concern.

In the remainder of this article, first we provide a brief introduction to permutation tests, primarily as applied to testing differences between two independent groups. Then we compare the indices just listed, paying particular attention to rank-order invariance. Finally, we explicitly compare the performance of permutation and parametric t tests, including cases with few subjects when assumptions required for parametric tests become more questionable than usual.

PERMUTATION TESTS

Permutation tests are not widely used, so some introduction is required. When conducting a parametric t test, users necessarily assume that the sampling distribution for their computed t follows the theoretically expected one, given that the null hypothesis is true, and on that basis assign a p value. Yet such p values become correct only asymptotically, as real conditions approach ideal assumptions. In contrast, permutation tests construct the relevant sampling distribution directly from the data at hand, thereby rendering the usual assumptions moot (this can be viewed as a Fisherian approach, as opposed to a classical or Neyman-Pearsonian one; see Camilli, 1990). As noted in Bakeman, Robinson, and Quera (1996), when discussing permutation tests applied to individual indices of sequential association:

> What are often called exact tests might better be called tests that yield exact, as opposed to approximate, p values. They are also called randomization tests (e.g., Bradley, 1968), although at least some experts (e.g., Edgington, 1987) reserve that term for tests involving data that result from random assignment of experimental units either to treatments or treatment times. *Permutation test* seems the more descriptive and more general term (Edgington, 1987), and is the term we use here. Like nonparametric tests, and unlike the common parametric z, t, and F tests, their derivation and application do not involve explicit assumptions about population distributions and parameters (Hays, 1963). And unlike the usual applications of non-parametric tests like those based on the chi-square distribution, they do not rely on asymptotic theory that is valid only if sample sizes are reasonably large and well balanced (Mehta & Patel, 1992).

A permutation test can be constructed for any statistic. For the case at hand, a log odds ratio or other statistic could be computed for each experimental unit and a t statistic for independent groups computed as usual; this is the observed value for t. However, instead of assuming that the observed t is distributed like the theoretical one tabled in statistics texts, its sampling distribution is constructed from the data at hand. Let N represent the number of scores and r the number in Group 1. Then the N subjects are divided into all possible groups of r and $N - r$ subjects, and a t statistic is computed for each different permutation. The exact probability of a result as extreme as the one observed is simply the proportion of these t statistics (absolute) greater than or equal to the magnitude of

the observed one (two-tailed test; a one-tailed test is confined to either the positive or negative direction).

This seems simple enough, and it may even seem surprising that permutation tests are not the norm. Yet few computer packages include them (although one package that does perform the appropriate computations for most common statistical tests is StatXact, Mehta & Patel, 1992; see also Lynch, Landis, & Localio, 1991). Moreover, although the appropriate permutations seem straightforward for a test involving two independent groups, they can become complex for other tests, both conceptually and computationally. Furthermore, the number of permutations can become huge. For two independent groups, the number is

$$\binom{N}{r} = \frac{N!}{r!\,(N-r)!},$$

which, with 20 subjects divided into two groups of 10, is 184,756 but, with just 40 subjects divided into two groups of 20, becomes over 138 billion.

Fortunately, there is a simple solution to the computing dilemmas created by the size and complexity of many problems. As noted by Edgington (1987) and Bakeman et al. (1996), among others, the sampling distribution for the test statistic can be based not on the complete set of permutations (an exact permutation test), but on a subsample instead (a sampled permutation test). For the present two-group example, subjects are ordered 1 through N, the first r subjects belong to Group 1, and the observed t is computed on this basis. Then the order is repeatedly shuffled (Castellan, 1992). After each shuffle (which yields a new permutation), the first r subjects are arbitrarily assigned to Group 1 and a new t statistic computed. After some large number of shuffles, the exact probability is estimated as the proportion of t statistics (absolute) greater than or equal to the magnitude of the observed t (two-tailed). But rather than constructing a sampling distribution based on all possible permutations, which might number in the billions, one based on a more reasonable number is constructed instead (Edgington, 1987, suggests that as few as 1,000 samples may be sufficient, but we routinely use 10,000).

One further simplification is possible. There is no need to compute a t statistic; the difference between the groups' means (i.e., the numerator used when computing t) can be used instead. This is, after all, the descriptive statistic of interest when testing for group differences. In usual practice, we divide the difference by its assumed standard error, thereby transforming it into a statistic whose theoretical distribution is known. This step is rendered unnecessary by permutation tests, which construct the sampling distribution from the data at hand. In fact, whenever test statistics are rank-order invariant over permutations, as t and the mean difference are, they produce equivalent permutation test results (Edgington, 1987), so, as a matter of convenience, it makes sense to use whichever test statistic requires the least computation. But now we return to a consideration of the indices of sequential association listed earlier, and ask whether there is any reason to prefer one over another when computing the test statistic used to test for group differences using sampled permutations.

INDICES OF ASSOCIATION

Odds Ratio and Log Odds Ratio

The central question posed here—which statistic and which test should be used for questions concerning group differences—was elegantly answered in a recent article by Wickens (1993). After considering several alternatives (including a partial chi-square derived from log-linear analysis, a chi-square statistic based on the Mantel-Haenszel estimate of the common log odds ratio, and an ad hoc F based on likelihood-ratio chi-squares), he concluded that for dichotomous classifications, a parametric t test on the log odds ratio was best. Wickens (1993) was particularly interested in comparing wholly log-linear (one dimension of a contingency table represents subjects) with parametric (scores derived for individual subjects subjected to subsequent t tests, etc.) approaches; the present paper continues his line of questioning, comparing parametric with permutation tests and considering indices such as Yule's Q and Wampold's transformed kappa in addition to the log odds ratio.

The log odds ratio is more familiar to epidemiologists than to most psychologists, so a brief review, beginning with the odds ratio, may be helpful. The odds ratio, as its name implies, is estimated by the ratio of a to b divided by the ratio of c to d,

$$\text{estimated odds ratio} = \frac{a/b}{c/d}, \qquad (1)$$

where a, b, c, and d refer to observed frequencies for the cells of a 2 × 2 table as noted earlier. (Notation varies, but for definitions in terms of population parameters, see Bishop, Fienberg, & Holland, 1975, and Wickens, 1993.) Multiplying numerator and divisor by d/c, this can also be expressed as

$$\text{estimated odds ratio} = \frac{ad}{bc}. \qquad (2)$$

Equation 2 is more common, although Equation 1 reflects the name and renders the concept more faithfully. Consider the following example:

	B	$\sim B$	
A	10	10	20
$\sim A$	20	60	80
	30	70	100

The odds for B after A are 1:1, whereas the odds for B after any other (non-A) event are 1:3; thus the odds ratio is 3. In other words, the odds for B's occurring after A are three times the odds for B's occurring after anything else. When the odds ratio is greater than 1 (and it can always

be made ≥1 by swapping rows), it has the merit, lacking in many indices, of a simple and concrete interpretation.

The odds ratio varies from 0 to infinity and equals 1 when the odds are the same for both rows (indicating no effect of the row classification). The natural logarithm (ln) of the odds ratio, which is estimated as

$$\text{estimated log odds ratio} = \ln\left(\frac{ad}{bc}\right), \tag{3}$$

extends from minus to plus infinity, equals 0 when there is no effect, and is more useful for inference (Wickens, 1993). However, Equation 3 estimates are biased. An estimate with less bias, which is also well defined when one of the cells is zero (recall that the log of zero is undefined), is obtained by adding ½ to each count:

$$y = \ln\frac{\left(a+\frac{1}{2}\right)\left(d+\frac{1}{2}\right)}{\left(c+\frac{1}{2}\right)\left(b+\frac{1}{2}\right)} \tag{4}$$

(Gart & Zweifel, 1967; cited in Wickens, 1993, Equation 8). As Wickens (1993) notes when recommending that the log odds ratio computed per Equation 4 be analyzed with a parametric *t* test, this procedure not only provides protection for a variety of hypotheses against the effects of inter subject variability when categorical observations are collected from each member of a group (or groups), it is also easy to describe, calculate, and present.

Yule's *Q*

Yule's *Q* is a transformation of the odds ratio designed to vary, not from zero to infinity with 1 indicating no effect, but from −1 to +1 with zero indicating no effect, just like the Pearson correlation. For that reason, many investigators find it more descriptively useful than the odds ratio. First, *c*/*d* is subtracted from the numerator so that Yule's *Q* is zero when *a*/*b* equals *c*/*d*. Then *a*/*b* is added to the denominator so that Yule's *Q* is +1 when *b* and/or *c* is zero and −1 when *a* and/or *d* is zero, as follows:

$$\text{Yule's } Q = \frac{\frac{a}{b}-\frac{c}{d}}{\frac{c}{d}+\frac{a}{b}} = \frac{\frac{ad-bc}{bd}}{\frac{bc+ad}{bd}} = \frac{ad-bc}{ad+bc}. \tag{5}$$

Yule's *Q* can also be expressed in terms of both the odds and log odds ratio. From Equation 1, and letting *x* represent the estimated odds ratio, $ad = x \times bc$. Substituting this expression for *ad* in Equation 5, dividing numerator and denominator by *bc*, and simplifying,

$$\text{Yule's } Q = \frac{x-1}{x+1}.$$

Similarly, from Equation 3, and letting *y* represent the estimated log odds ratio, $ad = e^y \times bc$. Again substituting this expression for *ad* in Equation 5, dividing numerator and denominator by *bc*, and simplifying,

$$\text{Yule's } Q = \frac{e^y-1}{e^y+1}.$$

Since Yule's *Q* can be expressed as a monotonically increasing function of both the odds and log odds ratio, these three indices are equivalent in the sense of rank ordering subjects the same way. Thus, for present purposes, comparisons with transformed kappa and phi need include only one of these three indices, and we selected Yule's *Q* because it shares the same −1 to +1 range with transformed kappa and phi.

Wampold's Transformed Kappa

A statistic similar to Yule's *Q* is Wampold's (1989, 1992) transformed kappa. Recognizing the need for a cellwise effect size statistic (one unaffected by total number of tallies), Wampold (1989) proposed a transformed kappa (which is somewhat different from the kappa often used to assess interobserver reliability, Cohen, 1960). Like all kappas, this one is based on the general formula

$$\kappa = \frac{x-m}{\max(x)-m}, \tag{6}$$

where *x* is an observed and *m* an expected score. However, one value is computed for each cell, whereas for Cohen's (1960) kappa a single value characterizes the entire table. Applied to transitional probabilities, or two-dimensional tables generally, and modifying Wampold's (1989) formulas to use tabular notation (x_{ij} represents a cell, x_{i+}, represents a row sum, etc.)—and reasoning that the maximum joint cell frequency, x_{ij}, can be no larger than whichever marginal total is smaller, row or column—kappa, when x_{ij} is greater than m_{ij}, is

$$\kappa'_{ij} = \frac{x_{ij}-m_{ij}}{\min(x_{i+},x_{+j})-m_{ij}}. \tag{7}$$

Its upper bound occurs when x_{ij} is the minimum of x_{i+} or x_{+j}, in which case transformed kappa is +1. However, when x_{ij} is less than m_{ij}, the lower bound for Equation 7 is not −1. Dividing it by an expression for its lower bound, and simplifying the quotient, yields

$$\kappa'_{ij} = \frac{x_{ij}-m_{ij}}{m_{ij}}, \tag{8}$$

which ensures that the lower bound for transformed kappa is −1 (see Wampold, 1989, 1992, for details and derivations). The prime added to kappa indicates transformed kappa as defined by Equation 7 when x_{ij} is greater than m_{ij} and by Equation 8 when x_{ij} is less than m_{ij}. Thus transformed kappa, like Yule's *Q*, ranges from −1 to +1, with zero indicating no association.

Transformed kappa was designed for two-dimensional tables of any size. Reformulated for a 2 × 2 table for comparability with the other statistics discussed here, the formula, when the observed is greater than the expected (Equation 7) and *a* + *b* is less than *a* + *c*, is

$$\kappa' = \frac{ad - bc}{ab + b^2 + ad + bd} = \frac{ad - bc}{(a + b)(b + d)}, \quad (9)$$

and when $a + c$ is less than $a + b$, is

$$\kappa' = \frac{ad - bc}{ac + c^2 + ad + cd} = \frac{ad - bc}{(a + c)(c + d)}. \quad (10)$$

Similarly, the formula for transformed kappa when the observed is less than the expected (Equation 8), reformulated for a 2 × 2 table, is

$$\kappa' = \frac{ad - bc}{a^2 + ac + ab + bc} = \frac{ad - bc}{(a + b)(a + c)}. \quad (11)$$

Squaring the denominator and enclosing it with a radical sign for the various expressions just given (Equations 9–11) emphasizes similarities with equations for phi and Yule's Q given in the next section (Equations 13–15). Derivations for Equations 9–11 are given in Appendix A.

The Phi Coefficient

One of the most common measures of association for 2 × 2 tables is the phi coefficient. This is simply the familiar Pearson correlation coefficient computed using coded data (Cohen & Cohen, 1983; Hays,1963). One definition for phi is

$$\phi = \frac{z}{\sqrt{N}}, \quad (12)$$

(where z is computed for the 2 × 2 table and hence equals $\sqrt{\chi^2}$. Thus, phi can be viewed as a z score corrected for sample size. Like Yule's Q and transformed kappa it varies from −1 to +1, with zero indicating no association. In terms of the four cells, phi is defined as

$$\phi = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}. \quad (13)$$

Multiplying and rearranging terms, this becomes

$$\phi = \frac{ad - bc}{\sqrt{(ac + bd + ad + bc)(ab + cd + ad + bc)}}. \quad (14)$$

If we now rewrite the expression for Yule's Q, first squaring the denominator of Equation 5 and then taking its square root,

$$\text{Yule's } Q = \frac{ad - bc}{\sqrt{(ad + bc)(ad + bc)}}, \quad (15)$$

the value of Yule's Q is not changed, but similarities and differences between phi and Yule's Q (Equations 14 and 15) are clarified, as we now discuss.

INDICES OF ASSOCIATION COMPARED

Multiplier and multiplicand in the denominator for Yule's Q (Equation 15) consist only of the sum of ad and bc, whereas multiplier and multiplicand in the phi denominator (Equation 14) add additional terms. Consequently, values for phi are always less than values for Yule's Q (unless b and c, or a and d, are both zero, in which case both Yule's Q and phi would be +1 and −1, respectively). In contrast, because the ad, bc sum is not contained in the denominator multipliers and multiplicands for transformed kappa (Equations 9–11), relations between values of Yule's Q and transformed kappa are more complex. However, Yule's Q and transformed kappa share a property that distinguishes them from phi. Both Yule's Q and transformed kappa are +1 when either b or c is zero and −1 when either a or d is zero (this is called weak perfect association; Reynolds, 1984), whereas phi is +1 only when both b and c are zero and −1 only when both a and d are zero (this is called strict perfect association). Thus, phi achieves its maximum value (absolute) only when row and column marginals are equal (Reynolds, 1984).

In addition to algebraic analysis of their formulas, these and other differences among the indices of association discussed here can be clarified further by examining selected numerical examples. We assumed that 100 transitions were tallied and that 20 began with A. For the first three examples, A was always followed by B (perfect positive association), and for the second three, A was never followed by B (perfect negative association), but the number of transitions ending in B varied from 20, to 30, to 40 for each set of three. The resulting 2 × 2 tables and associated statistics are shown in Figure 1. The leftmost table in Figure 1 represents strict perfect positive association (thus $\phi = 1$), the next two tables, weak perfect positive association, and the last three, weak perfect negative association. Accordingly, values for Yule's Q and transformed kappa are either plus or minus 1, whereas

	B	~B	B	~B	B	~B	B	~B	B	~B	B	~B
A	20	0	20	0	20	0	0	20	0	20	0	20
~A	0	80	10	70	20	60	20	60	30	50	40	40
Q	1		1		1		-1		-1		-1	
κ'	1		1		1		-1		-1		-1	
φ	1.00		.76		.61		-.25		-.33		-.41	

Figure 1. Six 2 × 2 tables and associated values for Yule's Q, transformed kappa, and phi; the first three tables indicate perfect positive association and the last three, perfect negative association.

	B	~B	B	~B	B	~B	B	~B	B	~B	B	~B
A	8	12	8	12	8	12	4	16	4	16	4	16
~A	12	68	22	58	32	48	16	64	26	54	36	44
Q	.58		.27		0		0		-.32		-.53	
κ'	.25		.14		0		0		-.33		-.50	
ϕ	.25		.11		0		0		-.11		-.20	

Figure 2. Six 2 × 2 tables and associated values for Yule's *Q*, transformed kappa, and phi; the first two indicate moderate positive association, the second two, no association, and the last two, moderate negative association.

values for phi vary, depending on the number of transitions that end in *B*. Thus, in the extreme, if *B* always followed *A* for all subjects, but frequencies of *A* and *B* were nearly equal for subjects in one group, whereas frequencies of *B* were double those of *A* for subjects in another group, analyzing phi would suggest group differences, whereas analyzing Yule's *Q* or transformed kappa would not.

One might conclude from the foregoing that Yule's *Q* or transformed kappa should be used in preference to phi. After all, if all *A*s are followed by *B*s, does it matter how many not-*A*s are also followed by *B*s? But this argument applies primarily when all cases are extreme (i.e., tables for most experimental units contain at least one zero or near-zero tally). Usually, few cells are zero and, quite properly, the index of association is affected by frequencies for both *A* and *B*. A second set of six tables, which have the same marginal totals as those in Figure 1 but represent moderate to no effects, are given in Figure 2. The first three tables on the left in Figure 2 are alike in that for each, 8 of 20 *A*s are followed by *B*s. But it matters whether there are 20 (first table), or 30 (second table), or 40 (third table) transitions that end in *B*, as reflected by the quite different values these tables generate for Yule's *Q*, transformed kappa, and phi.

The examples just presented consisted of tables whose total number of tallies was identical. Yet, when examining for group differences, often the number of tallies for the experimental units varies. Accordingly, we next generated 200 tables for which the number of transitions beginning with *A* was held constant at 20, but the number of *A*s followed by *B* was allowed to vary randomly between 0 and 20, the number ending with *B* between 20 and 40, and the total number of tallies between 60 and 100; thus, effects could vary from a perfect negative to a perfect positive association. Statistics were computed for each table, and values for phi and transformed kappa were plotted against Yule's *Q* (see Figure 3).

If values for phi and transformed kappa were rank-order invariant with respect to Yule's *Q* (which, as noted earlier, renders them equivalent for permutation tests), then the graphs in Figure 3 would show a smooth upward trajectory, whereas a jagged appearance indicates statistics that rank order cases differently. In this respect, phi performed considerably better than transformed kappa. Some small reversals in rank, especially for large nega-

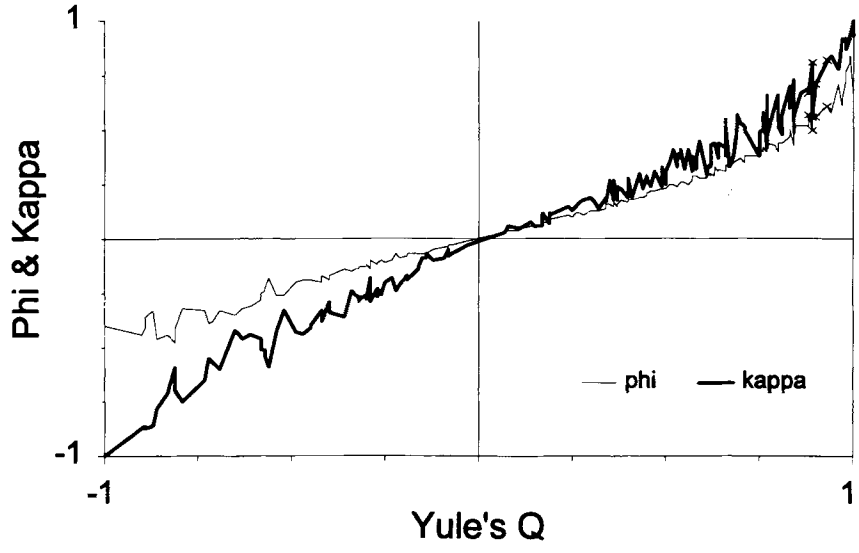

Figure 3. Yule's *Q* and associated phi (thin line) and transformed kappa (thick line) for 200 *A*/~*A* by *B*/~*B* tables generated as described in the text for which the observed value of *p*(*B*|*A*) varied from 0 to 1; tables associated with the points marked with ×s are given in Figure 4.

	B	~B	B	~B	B	~B	B	~B	B	~B
A	14	6	12	8	16	4	15	5	10	10
~A	21	44	17	56	21	26	18	29	10	50
Q	.66		.66		.66		.66		.67	
κ'	.49		.42		.55		.51		.33	
ϕ	.32		.33		.33		.34		.33	

Figure 4. Five 2 × 2 tables and associated values for Yule's *Q*, transformed kappa, and phi corresponding to points marked with ×s in Figure 3.

tive values, were seen with phi, whereas transformed kappa rank ordered far more cases differently. Often tables that gave similar values for Yule's *Q* and similar values for phi yielded divergent values for kappa. Five such tables, representing the five data points marked with ×s in the upper-right corner of Figure 3, are shown in Figure 4. An examination of Equations 9–11 for the 2 × 2 version of transformed kappa suggests why transformed kappa might yield such variable results: *c* is excluded from the denominator for Equation 9, *b* from the denominator for Equation 10, and *d* from the denominator for Equation 11. Thus, compared with phi and Yule's *Q*, transformed kappa may be viewed as less sufficient, where *sufficient* implies use of all the information available in the data and, along with efficiency and consistency, is usually listed as one of the attributes good estimators should possess.

A fourth desirable attribute is lack of bias, and here again transformed kappa may not perform as well as the other statistics discussed here. To investigate bias, we generated empirical sampling distributions for phi, transformed kappa, and Yule's *Q*. The distributions were based on 10,000 sequences, each 101 events long (so that 100 lag 1 transitions were tallied) and consisted of five different codes (including *A* and *B*). In line with the examples used earlier, the generating program set parameters representing probabilities for codes *A* and *B* to .2 and .3, respectively; probabilities for the remaining three codes were .2, .2, and .1. The degree of lag 0, lag 1 association was set to zero (i.e., each event in the sequence was selected randomly). Then transitions were tallied in *A*/~*A* by *B*/~*B* tables like those shown earlier, and phi, transformed kappa, and Yule's *Q* computed.

Distributions are shown in Figure 5. The mean for phi and Yule's *Q* was essentially zero (−.008 and .004, respectively), whereas the mean for transformed kappa was −.062, suggesting bias. Phi appears well formed and symmetric. As expected under these circumstances (i.e., unequal marginals for the 2 × 2 tables), the standard error (*SE*) for phi (.100) was less than the *SE*s for Yule's *Q* and transformed kappa (.272 and .226). Although not as well formed as phi, the distribution for Yule's *Q* appears reasonably symmetric, whereas transformed kappa evidences a markedly negative skew, with a negative mean and positive mode. On the basis of the evidence presented here, we would not recommend transformed kappa as an index of association for 2 × 2 tables, but recognize that it may have other uses (see Wampold, 1992). Again, on the basis of the evidence presented here, and other things being equal, when testing for group differences using permutation tests, we might have a slight preference for phi but suspect that it would yield essentially the same answers as Yule's Q (or the log odds ratio).

A final comment concerns cross influence. Sometimes sequential questions involve two interactants whose interaction is represented by two coded behavioral streams, not just one as in the examples presented here (see, e.g.,

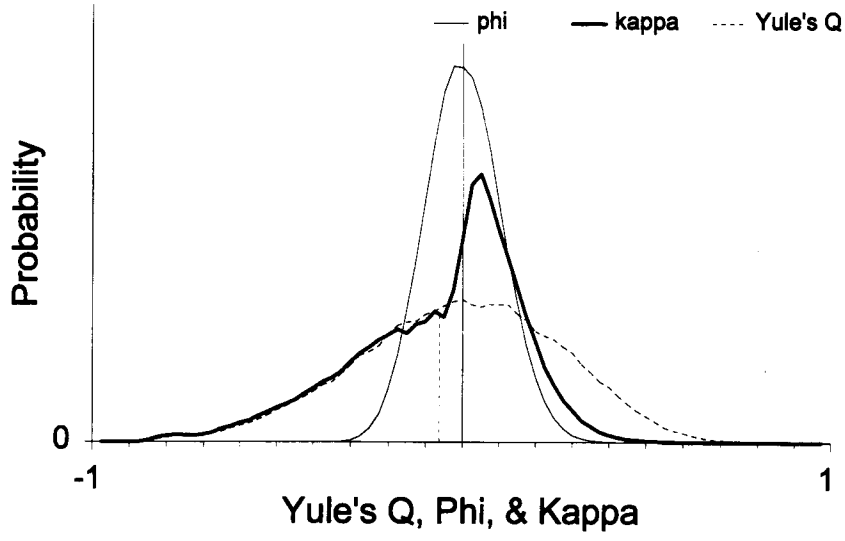

Figure 5. Sampling distributions for phi (thin line), transformed kappa (thick line), and Yule's *Q* (dashed line) based on 10,000 *A*/~*A* by *B*/~*B* tables for which the expected value of $p(B \mid A)$ was zero. The means for phi and Yule's *Q* were essentially zero; the mean for transformed kappa, represented by the vertical dotted line, was −0.062.

Allison & Liker,1982, and Faraone & Dorfman, 1987; for additional examples of situations more complex than those considered here, also see Budescu, 1984, and Iacobucci & Wasserman, 1988). In such cases, Dumas (1986), citing Tavaré and Altham (1983), argues that estimates of cross influence need to be corrected for autocorrelation within each stream, yet not all experts apply such corrections (e.g., Wickens, 1993), and in any event, the proposed correction is nil if autocorrelation is absent. Still, investigators who represent their data as two streams of coded data, and whose questions involve cross influence, should consult Dumas (1986).

PERMUTATION AND PARAMETRIC *t* TESTS COMPARED

Permutation tests seem attractive for reasons given earlier, but results might not differ dramatically from those produced by conventional parametric *t* tests. To compare these tests directly, several simulations were performed. For each simulation, the total number of subjects (N) and the number in the first group (r) were set. Values investigated for N were 10 (rs = 4 and 5), 15 (rs = 6 and 7), 20 (rs = 8 and 10), 40 (rs = 16 and 20), and 60 (rs = 24 and 30). Values for N were chosen to represent values that fall below as well as above usual guidelines regarding minimum sample size for parametric tests; thus 20 (10 subjects per cell) can be viewed as just satisfying a common rule of thumb. Values for r were chosen to represent the ideal (i.e., $N/2$) and slight deviations from that ideal. Consistent with tables presented earlier, we assumed, for each subject, that 20 transitions began with A and let the number of A to B transitions vary randomly between 0 and 20, the number ending with B between 20 and 40, and the total number between 60 and 100. For each subject, Yule's Q was computed for the $A/{\sim}A$ by $B/{\sim}B$ table. Then, for each set of N subjects—recall that each set was divided into two groups, one with r and one with $N - r$ subjects—a t statistic was computed and an exact probability based on 10,000 sampled permutations was estimated (as expected, probes showed identical results when phi instead of Yule's Q was used as an individual measure).

For each simulation, 100 sets of N subjects whose group means were somewhat different, as indicated by a t score between 1 and 3, were selected for further study. We wanted to compare asymptotic and permutation methods when the significance of mean differences would be most ambiguous, and this range contains one-tailed and two-tailed critical values for the values of N investigated here. Detailed results for two simulations are presented: the first, when N was 20 and r was 8, represents a case that just meets usual guidelines (Figure 6), and the second, when N was 10 and r was 4, represents a case that fails to meet usual guidelines (Figure 7).

With values of N as small as 20, and even with an r of 8 and not 10, the exact probabilities (technically, estimated exact probabilities since sampled permutation was used; see Bakeman et al., 1996) matched quite closely those expected for the parametric t (see Figure 6). The thick curved line represents asymptotic probabilities for the t scores listed on the abscissa, and most exact probabilities fall on or very near it. As a result, had we used an asymptotic instead of the more accurate permutation

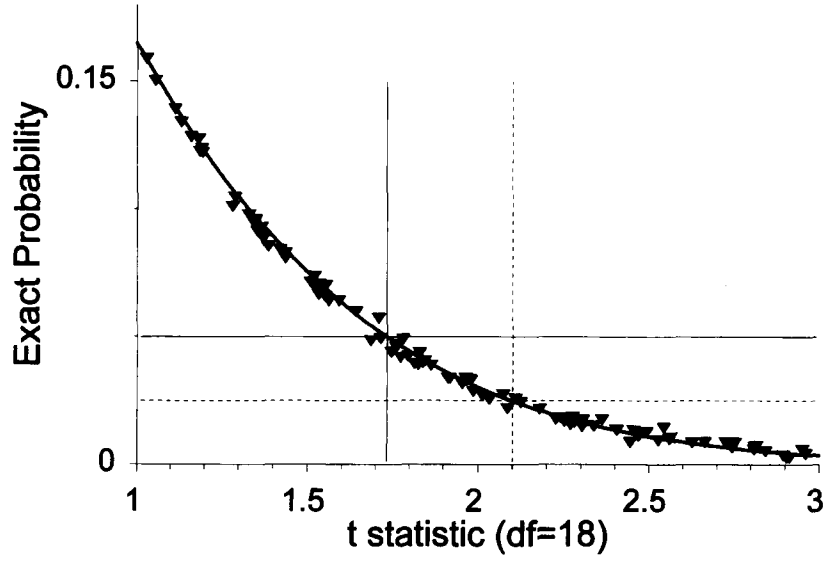

Figure 6. Estimated exact probabilities based on 10,000 sampled permutations and associated *t* scores for 100 sets of 20 subjects (8 in one group, 12 in the other) generated as described in the text. The thick curved line represents asymptotic *p* values for the associated *t* scores (*df* = 18). The thin and dotted horizontal lines represent exact probabilities of .05 and .025 (.05, two-tailed), respectively. The thin and dotted vertical lines represent *t* values of 1.734 (parametric *t* = .05, one-tailed) and 2.101 (parametric *t* = .05, two-tailed), respectively.

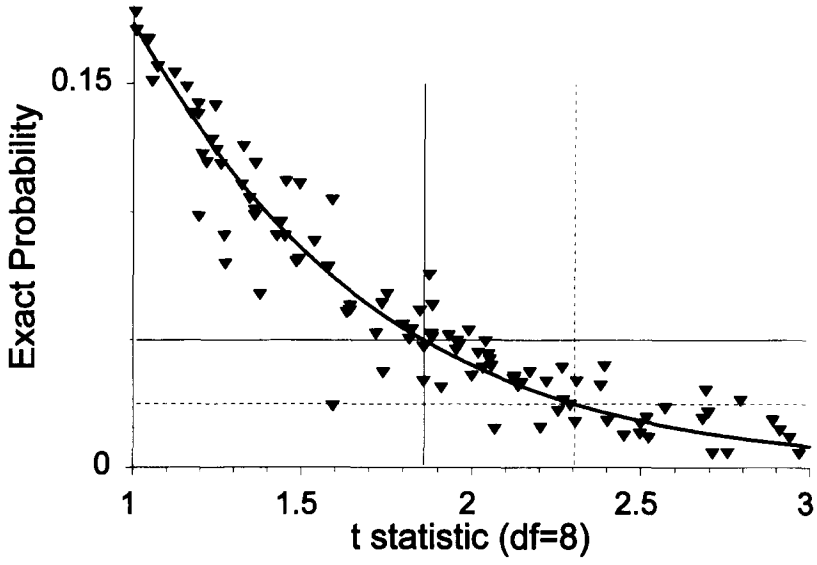

Figure 7. Exact probabilities based on 10,000 sampled permutations and associated *t* scores for 100 sets of 10 subjects (4 in one group, 6 in the other) generated as described in the text. The thick curved line represents asymptotic *p* values for the associated *t* scores (*df* = 8). The thin and dotted horizontal lines represent exact probabilities of .05 and .025 (.05, two-tailed), respectively. The thin and dotted vertical lines represent *t* values of 1.860 (parametric *t* = .05, one-tailed) and 2.306 (parametric *t* = .05, two-tailed), respectively.

test, one time we would have called a result significant when the exact test indicated no (two-tailed; parametric $t > 2.1$ yet exact probability > .025) and one other time we would not have called a result significant when the exact test indicated yes (parametric $t < 2.1$ yet exact probability < .025). But, in each case, the difference between asymptotic and exact probabilities was slight, as indicated by the small vertical distance between the asymptotic curve and the markers indicating exact probability (see Figure 6).

When N was 10, the exact probabilities fell further away from the asymptotic curve (see Figure 7). As a result, had we used an asymptotic instead of a permutation test, five times we would have called a result significant when the permutation test indicated no (two-tailed, parametric $t > 2.3$ yet exact probability > .025), and again five times we would not have called a result significant when the permutation test indicated yes (parametric $t < 2.3$ yet exact probability < .025). Differences between asymptotic and permutation probabilities were greater when N was 10 instead of 20. For example, one triangle in Figure 7 indicates an exact probability of .015, but its associated asymptotic probability is .036 ($t = 2.07$).

Given Figures 6 and 7, results of other simulations are easily described. When N was 10 and 20, probabilities derived from sampled permutation were somewhat closer to those for the parametric t when r was $N/2$; the results when N was 15 were more similar to those for 20 than 10; and when N was 20 or greater, even when r deviated from $N/2$, probabilities derived from sampled permutation were almost exactly the same as those for the parametric t.

SUMMARY AND RECOMMENDATIONS

When testing for group differences using sampled permutations, it does not much matter whether Yule's Q or phi is used, since both rank order cases essentially the same. Moreover, two monotonically increasing transformations of Yule's Q—the odds and log odds ratios—likewise give identical results. For a variety of reasons detailed here, these statistics are preferred over transformed kappa. Similarly, when subjects are 20 or more, and split reasonably evenly between two groups, it does not matter much whether sampled permutation or parametric t tests are used; both yield essentially the same result. However, when subjects are fewer than 20, or whenever there is any other reason to think that parametric assumptions may not be met, permutation tests are recommended. In any event, they are always the safer choice. Such tests are feasible, and a computer program that performs a two-independent-groups test is described in Appendix B.

REFERENCES

ALLISON, P. D., & LIKER, J. K. (1982). Analyzing sequential categorical data on dyadic interaction: A comment on Gottman. *Psychological Bulletin*, **91**, 393-403.

BAKEMAN, R., & GOTTMAN, J. M. (1986). *Observing interaction: An introduction to sequential analysis*. New York: Cambridge University Press.

BAKEMAN, R., & QUERA, V. (1992). SDIS: A sequential data interchange standard. *Behavior Research Methods, Instruments, & Computers*, **24**, 554-559.

BAKEMAN, R., & QUERA, V. (1995a). *Analyzing interaction: Sequential analysis with SDIS and GSEQ*. New York: Cambridge University Press.

BAKEMAN, R., & QUERA, V. (1995b). Log-linear approaches to lag-

sequential analysis when consecutive codes may and cannot repeat. *Psychological Bulletin*, **118**, 272-284.

BAKEMAN, R., ROBINSON, B. F., & QUERA, V. (1996). Testing sequential association: Estimating exact *p* values using sampled permutations. *Psychological Methods*, **1**, 4-15.

BISHOP, Y. M. M., FIENBERG, S. R., & HOLLAND, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

BRADLEY, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs, NJ: Prentice-Hall.

BUDESCU, D. V. (1984). Tests of lagged dominance in sequential dyadic interaction. *Psychological Bulletin*, **96**, 402-414.

CAMILLI, G. (1990). The test of homogeneity for 2 × 2 contingency tables: A review of and some personal opinions on the controversy. *Psychological Bulletin*, **108**, 135-145.

CASTELLAN, N. J., JR. (1979). The analysis of behavior sequences. In R. B. Cairns (Ed.), *The analysis of social interactions: Methods, issues, and illustrations* (pp. 81-116). Hillsdale, NJ: Erlbaum.

CASTELLAN, N. J., JR. (1992). Shuffling arrays: Appearances may be deceiving. *Behavior Research Methods, Instruments, & Computers*, **24**, 72-77.

COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educational & Psychological Measurement*, **20**, 37-46.

COHEN, J., & COHEN, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

CONGER, A. J., & WARD, D. G. (1984). Agreement among 2 × 2 agreement indices. *Educational & Psychological Measurement*, **44**, 301-314.

DUMAS, J. E. (1986). Controlling for autocorrelation in social interaction analysis. *Psychological Bulletin*, **100**, 125-127.

EDGINGTON, E. S. (1987). *Randomization tests* (2nd ed.). New York: Marcel Dekker.

FAGEN, R. M., & MANKOVICH, N. J. (1980). Two-act transitions, partitioned contingency tables, and the 'significant cells' problem. *Animal Behaviour*, **28**, 1017-1023.

FARAONE, S. V., & DORFMAN, D. D. (1987). Lag sequential analysis: Robust statistical methods. *Psychological Bulletin*, **101**, 312-323.

GART, J. J., & ZWEIFEL, J. R. (1967). On the bias of various estimators of the logit and its variance with application to quantile bioassay. *Biometrika*, **54**, 181-187.

HAYS, W. L. (1963). *Statistics* (1st ed.). NewYork: Holt, Rinehart, & Winston.

IACOBUCCI, D., & WASSERMAN, S. (1988). A general framework for the statistical analysis of sequential dyadic interaction data. *Psychological Bulletin*, **103**, 379-390.

LYNCH, J. C., LANDIS, J. R., & LOCALIO, A. R. (1991). StatXact. *American Statistician*, **45**, 151-154.

MEHTA, C., & PATEL, N. (1992). *StatXact: Statistical software for exact nonparametric inference*. Cambridge, MA: Cytel Software Corporation.

MORLEY, D. D. (1987). Revised lag sequential analysis. In M. L. McLaughlin (Ed.), *Communication yearbook* (Vol. 10, pp. 172-182). Beverly Hills: Sage.

REYNOLDS, H. T. (1984). *Analysis of nominal data*. Beverly Hills: Sage.

TAVARÉ, S., & ALTHAM, P. M. E. (1983). Serial dependence of observations leading to contingency tables, and corrections to chi-squared statistics. *Biometrika*, **70**, 139-144.

WAMPOLD, B. E. (1989). Kappa as a measure of pattern in sequential data. *Quality & Quantity*, **23**, 171-187.

WAMPOLD, B. E. (1992). The intensive examination of social interaction. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 93-131). Hillsdale, NJ: Erlbaum.

WICKENS, T. D. (1993). Analysis of contingency tables with between-subjects variability. *Psychological Bulletin*, **113**, 191-204.

APPENDIX A
Transformed Kappa Reformulated for 2 × 2 Tables

The formula for transformed kappa when the observed is greater than the expected (Equation 7), reformulated for a 2 × 2 table, is

$$\kappa' = \frac{a-(a+b)(a+c)/N}{\min\left[(a+b),(a+c)\right]-(a+b)(a+c)/N}$$

The a represents the observed and its expected value is $(a + b)(a + c)/N$. Replacing N with $a + b + c + d$ and simplifying, this becomes

$$\kappa' = \frac{a(a+b+c+d)-(a+b)(a+c)}{\min\left[(a+b),(a+c)\right](a+b+c+d)-(a+b)(a+c)}$$

Assume that $a + b$ is the minimum of $a + b$ and $a + c$. Then

$$\kappa' = \frac{a(a+b+c+d)-(a+b)(a+c)}{(a+b)(a+b+c+d)-(a+b)(a+c)}$$

$$= \frac{a^2 + ab + ac + ad - (a^2 + ac + ab + bc)}{a(a + b) + b(a + b) + c(a + b) + d(a + b) - (a^2 + ac + ab + bc)}$$

$$= \frac{a^2 + ab + ac + ad - a^2 - ac - ab - bc}{a^2 + ab + ab + b^2 + ac + bc + ad + bd - a^2 - ac - ab - bc}$$

$$= \frac{ad-bc}{ab+b^2+ad+bd} = \frac{ad-bc}{(a+b)(b+d)},$$

which is Equation 9 given earlier. Or, if $a + c$ is the minimum of $a + b$ and $a + c$, then

APPENDIX A (Continued)

$$\kappa' = \frac{a(a+b+c+d)-(a+b)(a+c)}{(a+c)(a+b+c+d)-(a+b)(a+c)}$$

$$= \frac{a^2 + ab + ac + ad - (a^2 + ac + ab + bc)}{a(a + c) + b(a + c) + c(a + c) + d(a + c) - (a^2 + ac + ab + bc)}$$

$$= \frac{a^2 + ab + ac + ad - a^2 - ac - ab - bc}{a^2 + ac + ab + bc + ac + c^2 + ad + cd - a^2 - ac - ab - bc}$$

$$= \frac{ad - bc}{ac + c^2 + ad + cd} = \frac{ad - bc}{(a + c)(c + d)},$$

which is Equation 10 given earlier.

The formula for transformed kappa when observed is less than expected (Equation 8), reformulated for a 2×2 table, is

$$\kappa' = \frac{a - (a + b)(a + c)/N}{(a + b)(a + c)/N}.$$

Again replacing N with $a + b + c + d$ and simplifying, this becomes

$$\kappa' = \frac{a(a+b+c+d)-(a+b)(a+c)}{(a+b)(a+c)}$$

$$= \frac{a^2+ab+ac+ad-a^2-ac-ab-bc}{a^2+ac+ab+bc}$$

$$= \frac{ad-bc}{a^2+ac+ab+bc} = \frac{ad-bc}{(a+b)(a+c)},$$

which is Equation 11 given earlier.

APPENDIX B
PGD, a Permutation Program for Testing Group Differences

A computer program for testing differences among groups using permutation tests has been developed. This program reads data from an ASCII file containing scores for individuals, or units of analysis, in different groups, and provides an estimate of the exact p value of the one-way analysis of variance Snedecor's F test. If requested, it also computes estimates of two-tailed exact p values for all possible independent Student's t tests among groups (if there are only two groups, p values for F and t tests are identical). PGD computes p values by permuting the observed data repeatedly and computing the proportion of permutations that yield statistics (F and t) that are greater than or equal to the observed ones. All the data are shuffled in order to compute p values for F tests, whereas data for groups being compared are shuffled independently from the other groups when multiple comparisons among groups are performed. Confidence intervals for the p values are also computed on the basis of blocks of permutations. Defaults used by PGD are: (1) Observed data are permuted or shuffled 1,000 times, using Castellan's (1992) shuffling algorithm. (2) The procedure is repeated 10 times (i.e., a total of 10,000 times, 10 blocks of 1,000 permutations each) in order to compute mean p values and 95% confidence intervals. (3) Multiple-comparison tests are not performed.

Defaults can be changed by users in several ways by invoking PGD with some optional arguments in the DOS command line: (1) Number of permutations, up to a maximum of 5,000 per block. (2) Number of blocks, up to a maximum of 50 (i.e., total maximum number of permutations can be 250,000). (3) 99% confidence intervals can be requested. (4) Multiple-comparisons procedure among groups can be requested. (5) Printout of first shuffled data can be requested, up to a maximum of 50.

APPENDIX B (Continued)

When multiple comparisons are specified, PGD reports both the p values and the corresponding adjusted values resulting from using Fisher's *modified least significant difference* procedure, which takes into account the fact that multiple t tests are being performed. This procedure consists of multiplying each p value by the number of comparisons in order to get the final adjusted significance values (Edgington, 1987, pp. 85–88).

PGD is written in Borland C and runs in an MS-DOS environment. It is available from the authors. To receive a copy of the executable file, mail a formatted 3.5-in. floppy disk with a self-addressed return mailer to either author (stamped if within USA or Spain). Enclose a note stating that you will use the program only for noncommercial purposes, that you will not give copies to others, and that you understand that it is not guaranteed to be free of error.

(Manuscript received August 25, 1994;
revision accepted for publication June 6, 1995.)