Behavior Research Methods, Instruments, & Computers
2001, 33 (3), 331-338

Comparing sequential associations within a single dyad

PAUL J. YODER
Vanderbilt University, Nashville, Tennessee

PETER BRUCE
Resampling Stats, Arlington, Virginia

and

JON TAPP
Vanderbilt University, Nashville, Tennessee

We present a new application of sampled permutation testing to examine whether two sequential associations are different within a single dyad (e.g., a teacher and a student). A Monte Carlo simulation with the same (i.e., 100 vs. 100) or a different (100 vs. 400) number of event pairs was used to simulate designs that use time-based (typically producing equal-length comparisons) and event-based (typically producing different-length comparisons) data, respectively. For these pairs of simulated data streams, we compared the Type I error rates and the kappa for agreement on significance decisions, using the sampled permutation tests and the more traditional asymptotic log linear analysis. The results provide the first evidence relevant to evaluating the accuracy of log linear analysis and sampled permutation testing for the purpose of comparing sequential associations within a single dyad.

Is the sequential association between one possible trigger event and a particular target event stronger in one context than in another context? This paper presents a new application of sampled permutation tests (i.e., *shuffle-the-cells* sampled permutation test) to examine such questions within a single subject. A sampled permutation test uses repeated sampling of shuffling of the sequences of the data at hand to create an empirical probability distribution of the observed difference in sequential associations (Good, 2000).

Note that we are concerned with determining the statistical significance of the difference among associations between a single subject's behaviors. In other words, we are asking (for example) whether behavior A follows behavior B more in one dimension of comparison than in another (e.g., different contexts) for a single subject and whether this difference between dimensions might have occurred by chance. In this application of significance testing, we are not concerned with behavior tendencies in a population of people (Hayes, 1996).

There are many situations in which one wishes to test whether one sequential association is greater than another *within a single dyad.* Doing so maximizes the clinical value of the results of sequential analyses because treatments designed to alter the occurrence of antecedent events (i.e., the first behavior in a pair of behaviors) to reduce or increase the target events (i.e., the second behavior in a pair of behavior) are implemented at the individual level. To the clinical scientist, there is great value in being able to test one's research question within every subject. When one tests a hypothesis within a single subject or dyad, one is testing whether an observed data pattern can be characterized as random for that subject, not whether the observed data is characteristic of a population of subjects (Hayes, 1996). External validity is addressed through repeated replications. Potentially, such a strategy allows as many replications of the predicted results as there are dyads in one's sample.

BRIEF EXAMPLE AND OVERVIEW OF SEQUENTIAL ANALYSIS

One example of a research question involving the comparison of sequential associations is as follows: "Is the extent to which student topic-continuing utterances follow teacher instruction greater in class than on the playground." Note that our real interest is in whether teacher instruction may elicit topic-continuing utterances. Also note that, on the playground, "teacher instruction" might be less likely to be followed by "student topic-continuing utterance" than in the classroom simply because the latter behavior does not happen often on the playground, and not because the association between the two is weaker. Therefore, we want to control for the different base rates

This article was written while P.J.Y. and J.T. were supported by Grant NICHD P30HD15052. We gratefully acknowledge the support of the Kennedy Center. Correspondence concerning this article should be addressed to P. J. Yoder, Kennedy Center, Vanderbilt University, Nashville, TN 37215 (e-mail: paul.yoder@vanderbilt.edu).

of the behaviors when comparing the strength of the association between the two behaviors across contexts.

A more precise restatement of the example research question is "Controlling for differences in the frequency of occurrence in the relevant behaviors, is the extent to which *student topic-continuing utterance* follows *teacher instruction* greater in class than on the playground?" A sequential association is the extent to which one behavior follows another after controlling for chance sequential occurrences (Bakeman & Gottman, 1997). To address research questions involving the comparison of sequential associations in different contexts within a single dyad, it is useful to compute an index of sequential association between the two critical behaviors within each context. Ideally, the index quantifying this close temporal association controls for the simple probability of the two behaviors of interest (i.e., their base rate; Bakeman, McArthur, & Quera, 1996).

To aid in the computation of such an index, organizing the sequence of behavior into a 2 × 2 table is helpful. Such a table can be constructed to organize any sequence of events. In the 2 × 2 contingency table, we organize all pairs of behaviors into a mutually exclusive and exhaustive table with two rows (first behavior in the pair) and two columns (second behavior in the pair). The top row is labeled for the proposed antecedent (e.g., *teacher instruction*) and the first column is labeled for the proposed target (e.g., *student topic-continuing utterances*). To refer to the data in each cell of the 2 × 2 table, we assign the following labels: Cell A (top left) = antecedent preceding target behavior, Cell B (top right) = antecedent preceding nontarget and nonantecedent behavior, Cell C (bottom left) = nonantecedent and nontarget behavior preceding target behavior, and Cell D (bottom right) = nonantecedent and nontarget behavior preceding other instances of nonantecedent and nontarget behavior (Figure 1). In our present example, the 2 × 2 tables represent what occurs in a particular context, but there are other applications of sequential analysis that compare 2 × 2 tables (Bakeman & Gottman, 1997; Yoder & Feurer, 2000; Yoder, Short-Meyerson, & Tapp, 2000).

We have selected Yule's Q as our index of sequential association because it controls for the base rate of the antecedent and target behavior and is not influenced by the total number of events in the analyses (Bakeman, McArthur, & Quera, 1996; Yoder & Feurer, 2000). However, we see no reason why the results of the present Monte Carlo would not generalize to other indices of sequential association that control for the base rates of the target and antecedent events and fit the requirements of the research question and context. Yule's Q is essentially the odds ratio scaled to a range between -1 and 1 (Bakeman, McArthur, & Quera, 1996). Using the cell labels for the 2 × 2 table, Yule's Q is computed as follows:

$$\text{Yule's } Q = (\mathrm{AD} - \mathrm{BC})/(\mathrm{AD} + \mathrm{BC}).$$

To measure the difference in sequential association between one context and another, we compute the difference score for the Yule's Q scores from the two 2 × 2 tables representing the sequential associations to be compared.

The null hypothesis using Yule's Q scores is as follows:

$$Q1 = Q2 \text{ or } Q1 - Q2 = 0.$$

Note that we are not testing the null hypothesis that the sequential association between teacher instruction and student topic-continuing utterances is at the chance level.

Sampled Permutation Tests

Permutation tests examine all possible permutations of the data under the null model, recalculate the test statistic for each permutation, and count what proportion of the permutations produced a test statistic as extreme as that observed. They guarantee preservation of Type I error rates at the nominal level of the test without making assumptions about the shape of sampling distributions, sample size, or marginal distributions (Agresti, 1990; Blair & Karniski, 1993; Bradley, 1968; Cervone, 1985; Edgington, 1964; Kratochwill et al., 1974; Marascuilo & McSweeney, 1977; Pitman, 1937; van den Brink, 1988; Wampold & Worsham, 1986).

Sampled permutation tests, also called Monte Carlo permutation tests, take a random sample of permutations rather that treating all permutations exhaustively. As the number of Monte Carlo simulations approaches infinity, the Monte Carlo p value approaches the exhaustive per-

		Second Behavior		
		Behavior B	Other	Total
First behavior	Behavior A	A	B	
	Other	C	D	
	Total			

Figure 1. The 2 × 2 contingency table and cell labels.

Table 1
Transitional Probabilities of the Simulated Behavior Streams

	Second Behavior		
First Behavior	Antecedent	Target	Other
	Population Yule's $Q = .50$		
Antecedent	.09	.26	.64
Target	.42	.15	.43
Other	.32	.09	.59
	Population Yule's $Q = .73$		
Antecedent	.09	.45	.46
Target	.42	.15	.43
Other	.32	.09	.59

Note—To populate the 2 × 2 contingency tables used to calculate Yule's Q, three behavior types are of interest (antecedent, target, and behaviors that are neither antecedent nor target); hence, nine (i.e., 3 × 3) transitional probabilities are required to determine the probability of the behavior types following one another.

mutation p value. Ten thousand permutations generally produce a very good approximation to the exact p value (Edgington, 1995).

Sampled permutation tests provide an analysis strategy that is well suited to significance testing within a single subject or dyad. Permutation tests have been used to test significance of test statistics within a single case by many investigators (Edgington, 1975, 1980a, 1980b, 1996; Ferron & Onghena, 1996; Ferron & Ware, 1994; Levin, Marascuilo, & Hubert, 1978; McLeod, Taylor, Cohen, & Cullen, 1986; Onghena & Van Damme, 1994; Yoder, Klee, Hooshyar, & Schaffer, 1997). Bakeman, Robinson, and Quera (1996) provided an example of how sampled permutation tests can be used to test the significance of a single sequential association within a single subject.

Shuffle-the-Cell Permutation Tests

The p value of the observed difference in the sequential associations is calculated by determining how often the permuted (shuffled) *difference* in Yule's Q is as extreme as the observed difference between the two contexts:

$$\text{(frequency that } |\text{post-shuffle } Q1 - \text{post-shuffle } Q2| >= |\text{observed } Q1 - \text{observed } Q2|)/ \text{ number of permutations.}$$

We take the absolute value of observed and permuted difference scores to implement a two-tailed test.

Because our null model is that the sequential association in one dimension to be compared (e.g., context) is the same as that in the other, we first combine together the classified behavior pairs that make up the 2 × 2 contingency tables from which Yule's Q is calculated for both contexts. Then we shuffle and redivide them into two samples of the original sizes.

Specifically, recall that earlier we labeled each pair of behaviors according to the cell into which it falls, using the cell labels A, B, C, D. Now, pairs from Table 1 and Table 2 are combined together before permutation (i.e., shuffling the behavior pairs across tables). This step embodies the null hypothesis that the sequential associations are the same, regardless of context.

For example, one way to represent 10 pairs of behaviors for the *shuffle-the-cell* sampled permutation test is A B B C C D D D D D. Using the same cell labels as were used for the first table, the data from a second 2 × 2 table containing another 10 pairs of behaviors are represented as A A A A B C C D D D. We can test the null hypothesis that the magnitude of the sequential association is equal across contexts by using the following permutation procedure:

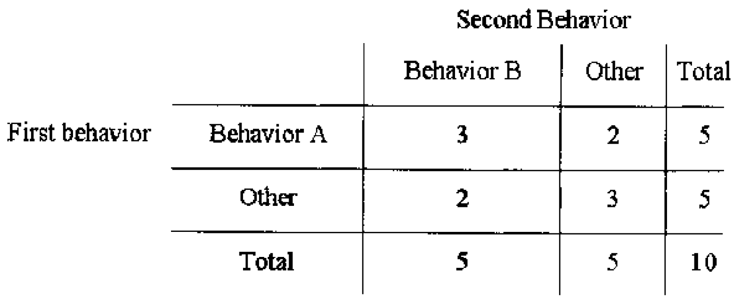
1. Join the two data lists that represent our 2 × 2 table cell tallies of behavior pairs together (i.e., all the As, Bs, Cs, and Ds).
2. Shuffle the combined data list and split it into two permuted samples of the same size as the observed samples.
3. Create the two new 2 × 2 tables by counting the number of As, Bs, Cs, and Ds in the two permuted samples.
4. Compute the two indices of sequential analysis for the two new 2 × 2 tables.
5. Take the difference between these indices from permuted samples.
6. Repeat this permutation procedure 10,000 times to create the permutation distribution of Yule's Q difference scores against which the observed Yule's Q difference score can be tested.

For example, the combined data list for our two observed 2 × 2 tables is as follows: A B B C C D D D D D A A A A B C C D D D. Using a random process, we shuffle the combined data list. An example of the results of shuffling the combined data list might look like this: B D A D B C C D A A C D C D A B D D A D. Next, we create the two new 2 × 2 tables. To construct the new tables, we select 10 values from the shuffled combined data list (i.e., the total number of behavior pairs in the first 2 × 2 table). For example, the first permuted sample could be as follows: B D A D B C C D A A. Next, we construct a 2 × 2 table on the basis of this data list. The 2 × 2 table for permuted data representing the first context is shown in Figure 2. We repeat this process to re-

Table 2
Type I Error Rates of the Two Significance Test Methods Examining the Difference Between Sequential Associations in Pairs of Behavior Streams

Statistical Analysis Method	Type I Error Rate*
Population Yule's $Q = .50, N = 100$	
Shuffle-the-cell permutation	.047
Log linear	.056
Population Yule's $Q = .73, N = 100$	
Shuffle-the-cell permutation	.059
Log linear	.066

*Proportion of 1,000 Monte Carlo trials with significant p values; $\alpha = .05$.



		Second Behavior		
		Behavior B	Other	Total
First behavior	Behavior A	3	2	5
	Other	2	3	5
	Total	5	5	10

Figure 2. The permuted table for Context 1.

construct the 2 × 2 table for a permuted equivalent for the second context. Because we sampled without replacement, the 10 values for this 2 × 2 table are the remaining 10 values. The second permuted table is shown in Figure 3. Yule's *Q* scores are computed for each table, and the difference score is computed. This process is then repeated 10,000 times. We recorded how many of the permutations produced a difference in Yule's *Q* scores as extreme as the observed difference in Yule's *Q* scores. Of course, in executing this test, one would not display the 2 × 2 tables at each iteration; one would simply use the underlying data to compute and record the difference in Yule's *Q*.

Log Linear Analysis To Compare Sequential Associations Within an Individual

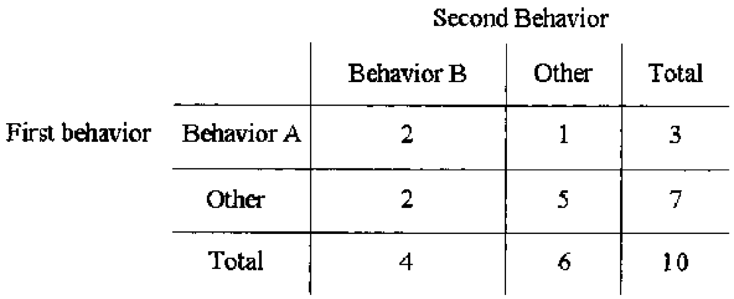
In the currently published studies using sequential analysis, the most frequently used method to compare sequential associations within a single subject is log linear analysis (e.g., Krokoff, Gottman, & Roy, 1988). To compare two sequential associations using log linear analysis, one tests whether the three-way interaction between the first behavior, the second behavior, and the dimension being compared across (e.g., context) is significant (Bakeman & Quera, 1995).

Applied to the comparison of two sequential associations from different contexts within a single case, a 2 × 2 × 2 table is constructed (first behavior × second behavior × context). The maximum likelihood method is used to estimate the expected contingency table cell values. A likelihood ratio chi-square statistic is the test statistic that is tested against the chi-square distribution (Knoke & Burke, 1980). More detailed information for understanding and interpreting log linear analysis in comparing sequential associations can be found in Bakeman and Quera (1995).

The log linear method relies on the use of large sample theory and a theoretical sampling distribution to approximate the exact sampling distribution of the test statistic in question. When sample sizes are large and adequately balanced, the chi-square approximation serves well in the analysis of contingency table data. However, when sample sizes are small or when the data are not well balanced (e.g., when one event occurs relatively infrequently), the approximation is less accurate (Agresti, 1990; Edgington, 1995). Additionally, there is currently no study that has tested the validity of the asymptotic *p* value from log linear analyses when applied to testing the difference in sequential associations with a single dyad. Next, we present a series of Monte Carlo studies that provide such information.

The Monte Carlo Simulations

We used Monte Carlo simulations to compare the results of the log linear method and the shuffle-the-cell sampled permutation method when sampling from a



		Second Behavior		
		Behavior B	Other	Total
First behavior	Behavior A	2	1	3
	Other	2	5	7
	Total	4	6	10

Figure 3. The permuted table for Context 2.

population with known characteristics. In each set of Monte Carlo simulations, we randomly generated many pairs of data streams from the same algorithm to produce a null situation. Each pair of data streams is analogous to a pair of behavior streams observed in two different contexts. To ensure that the characteristics of the generated behavior streams were examples of what could occur in nature, we wrote a computer program using the transitional probabilities from two actual adult–child conversations. We selected conversations that yielded a moderate positive sequential association (Yule's $Q = .50$) and a high positive sequential association (Yule's $Q = .73$). These conversations were used to identify the transitional probabilities that would yield these moderate and high Yule's Q scores. In each of the generated behavior streams, there were three types of behaviors: a proposed antecedent behavior, a proposed target behavior, and one other type of behavior. The probability of any of the three behaviors' being generated as the first behavior in the stream was .33. The subsequent behaviors were generated randomly, but with the constraints of the transitional probabilities (Table 1). The computer program was written to reject and replace behavior streams with undefined Yule's Q scores (i.e., zero cells in either Cell A or D *and* in Cell B or C).

All but the last of the Monte Carlo simulations used behavior streams with 100 events (see Tables 2 and 3 for details). Admittedly, 100 events is a small sample for sequential analyses (Bakeman & Gottman, 1997). We began our examination with such a small sample size because asymptotic tests such as log linear analysis tend to produce too many Type I errors with small samples (Tansey, White, Long, & Smith, 1996). If the significance tests produce accurate and similar Type I error rates to the shuffle-the-cell permutation test under such small sample conditions, larger samples should produce even more accurate Type I error rates.

Two sets of Monte Carlo simulations were conducted. In the first, we sought to compare the Type I error rate and kappa on agreement of significance decisions of log linear and shuffle-the-cell for equal length behavior streams ($N = 100$) for both levels of sequential association (i.e., .50 and .73). The prediction was that the shuffle-the-cell method would have a more conservative and more accurate Type I error than would log linear analysis.

In the second series of simulations, we compared the Type I error rates and agreement on significance decisions for the log linear and shuffle-the-cell methods when comparing behavior sequences of different lengths. One member of the pair had 100 events, and the other had 400 events. The need to compare different-length behavior streams is common in the real world when event-based sequential analyses are conducted. In the second simulation, we used the algorithm designed to produce a mean Yule's Q of .73 to generate the behavior streams.

Both significance test methods (i.e., log linear analysis and shuffle-the-cell sample permutation tests) were applied to the same pairs of behavior streams to allow comparison of Type I error rates and to compute the extent to which the significance tests agreed on significance decisions. Statistical significance was assessed by using a two-tailed test, with alpha level set at .05. The Type I error rate was the proportion of generated behavior pairs that had a statistically significant difference in sequential associations, even though there was no difference in the population. Kappa (Cohen, 1960) was the index of agreement. We generated 1,000 Monte Carlo trials to estimate Type I error rates and kappa.

Table 3
Kappa and Type I Error Rates for Log Linear and Shuffle-the-Cell Methods When Comparing a Behavior Stream With 100 Events With Another With 400 Events

Kappa on significance decisions	.58
Type I error rate	
Log linear	.052
Shuffle-the-cell permutation tests	.048

RESULTS

In the first set of simulations, no generated behavior streams had to be rejected due to undefined Yule's Q scores. The algorithms were successful in producing behavior streams with the desired characteristics. The mean of the sampling distribution for both of the 2×2 tables in the first set of behavior stream pairs was .50 ($SD = .21$), and the mean for both of the 2×2 tables in the second set of behavior stream pairs was .73 ($SD = .38$). The mean difference between Q scores was not significantly different from zero in either data set (mean difference for Set 1 $= -.008$, $SD = .35$; mean difference for Set 2 $= -.001$; $SD = .18$). The mean transitional probabilities for both sets of 1,000 behavior streams matched those in Table 1 to within less than a percentage point.

Equal-Length Condition

Table 2 presents the Type I error rates for shuffle-the-cell and log linear analyses for comparing the sequential associations of the equal-length behavior streams. The Type I error rates were very similar and quite close to .05, the expected Type I error rate. Additionally, the kappas for agreement on significance decisions between the two significance-testing methods were over .90 in both cases.

Different-Length Condition

None of the generated behavior streams yielded undefined Yule's Q scores. The mean difference between Q scores was $-.007$ ($SD = .14$; $t = -1.78$, $p = .087$). Therefore, the null condition of zero difference in sequential associations was generated. The results for this last simulation are shown in Table 3. It should be noted that although the Type I error rates were similar and accurate, the degree of agreement on significance decisions was poor.

When significance testing methods disagree, we need to know which to use. One way to do so is to favor the method in which the data best match the assumptions of the analysis method. One of the assumptions of the asymptotic log linear analysis method is that the data are sampled from a Poisson distribution (Tansey et al., 1996). We tested the distribution of the truncated form of the odds ratio (i.e., rounded to the nearest integer) instead of their Yule's *Q* scores because the former fits the possible values of the Poisson distribution and Yule's *Q* is a linear transformation of the odds ratio (Bakeman, McArthur, & Quera, 1996). Poisson distributions only apply to integer data.

Therefore, we tested whether the 2 × 2 table cell values and the truncated form of the resulting odds ratios for the two 2 × 2 tables were Poisson distributed from the entire sample and two subsamples on which the significance tests disagreed. The truncated odds ratio from the entire sample for both 2 × 2 tables were not Poisson distributed ($Z = 2.94$, $p < .0001$; $Z = 7.01$, $p < .0001$, for Tables 1 and 2, respectively). This occurred because the cell values for the B and D cells from the entire sample for both 2 × 2 tables were not Poisson distributed ($Z = 1.4$, $p = .04$; $Z = 5.5$, $p < .001$; $Z = 1.6$, $p = .01$; $Z = 5.4$, $p < .001$, for B and D cells in Table 1 and Table 2, respectively). Additionally, the D cell from the longer behavior stream was not Poisson distributed ($Z = 1.37$, $p = .05$) in the subsample for which only the shuffle-the-cell method was significant. All other cells and odds ratios from the entire sample and disagreement subsamples were Poisson distributed.

DISCUSSION

When behavior streams to be compared were the same length, the Type I error rates of nearly .05 and the high agreement on significance decisions for the shuffle-the-cell and log linear methods support the conclusion that both are appropriate to test the difference in sequential associations. When sequential associations from behavior streams of very different lengths are compared, the Type I error rates were approximately correct, but the shuffle-the-cell and log linear methods disagreed frequently in regard to the statistical significance of the difference between particular pairs of sequential associations.

When the tests disagree, we want to know which of the tests to believe. The difference in significance decisions may have been due to the different statistic used by the two methods or to the method of estimating the probability of the observed difference in sequential associations. Log linear analysis uses the likelihood ratio chi square, whereas this application of shuffle-the-cell uses Yule's *Q*. Log linear analysis uses the chi-square distribution, and the shuffle-the-cell method uses an empirical distribution to estimate the probability of observed differences. Several of the cell values from the 2 × 2 tables violated the log linear test's assumption that the data be Poisson distributed. This leads us to hypothesize that when such violations occur and the data streams are of different lengths, the test methods will produce different results. Future work is needed to test this hypothesis and to determine other conditions under which the test methods produce different results. In the meantime, when comparing behavior streams of different lengths, we recommend using the shuffle-the-cell method because it does not require that data be sampled from a Poisson-distributed population.

In the situations discussed in this paper, analysis units (i.e., what is being permuted) are assumed to be "exchangeable" under the null hypothesis (Good, 2000). Here, the analysis units are pairs of behaviors. The exchangeability assumption underlying our null model (equal associations across contexts) is that there is a common "supply" of such pairs and their assignment to Context 1 or Context 2 is random. This assumption will be met in most real-world applications. Our example situations meet this condition because any type of behavior pair (represented by A, B, C, or D) could occur in either context. Therefore, shuffling behavior pairs across tables modeled the null situation without violating what could happen in reality by chance.

As with most significance tests, permutation tests assume that behavior pairs occur independently of one another. That is, it is assumed that they do not influence the subsequent occurrence of other behavior pairs under the null situation (Good, 2000; Hayes, 1996). The Monte Carlo simulation generated behavior streams with independently sequenced behavior pairs. Recall that the algorithm that generated behavior streams used only the transitional probability of one behavior to the next behavior. Therefore, we see no reason for why the data in the simulated pairs of behavior streams would violate the assumption of independent behavior pairs in any substantively important way. Indeed, the nearly accurate Type I error rates support this conclusion. It should be noted that any statistical analysis testing the difference in sequential association across contexts assumes that analysis units are sampled independently under the null condition (Good, 2000).

It is one thing to note that the assumption of exchangeability has not been violated in a simulation and another to claim this in reality. It is the latter that affects whether our *p* values are accurate in the real world. Because both asymptotic and permutation methods assume independent analysis units, we want to know the consequences of violating this assumption. This is particularly important because many have claimed that permutation tests do not assume independent analysis units (e.g., Edgington, 1980a). In other words, we want to know whether permutation tests are robust despite violations of the assumption of exchangeability.

As applied to comparing sequential associations, we do not know the answer. It has been shown that both asymptotic tests and permutation tests of the significance of correlation coefficients produce noteworthy elevations in Type I error rates when *some* subjects' data within *both* variables are dependent (Hayes, 1996). This is what Hayes (1996) called *mixed nonindependence* for both variables. Interestingly, nonindependence that is uniform across subjects (e.g., all subjects are in the same classroom) or that occurred in only one of the variables of interest (e.g., being in the same class affected students' knowledge of a subject, but not their IQ levels) did not affect the Type I error rate appreciably (Hayes, 1996). Unfortunately, we do not yet have empirical demonstrations of the conditions under which dependence of behavior pairs are problematic for tests of significance of the difference between sequential associations. Until such research is available, we recommend that one expect the conditions associated with elevated Type I error rates for tests of the difference between sequential associations across contexts to be the same as those found for the significance of correlation coefficients (Hayes, 1996). That is, dependence between some, but not all behavior pairs, in both contexts (i.e., 2 × 2 tables) in the null situation (sequential associations are the same across contexts) are expected to produce particularly elevated Type I error rates. This is expected to be the case for both log linear and shuffle-the-cell permutation tests.

At present, we recommend that the investigator use theory and clinical knowledge to estimate the extent to which their data fit this pattern. If the data logically fit this pattern, a sufficient number of behavior pairs could be discarded from the analysis to "break the dependence" between the sequence of behavior pairs. The necessity for and number of the discarded behavior pairs must be determined by the investigator and will vary across research questions and data sets.

In summary, testing the difference in sequential associations within a single dyad may have much clinical value. Clinical decisions are made with individuals, not groups of subjects, in mind. Both log linear and shuffle-the-cell permutation tests offer viable significance tests in many situations. The test of choice depends on whether the behavior streams to be compared are of the same length. Proper caution should be exercised in interpreting the results of tests in which the data violate the assumption of independence of analysis units. This assumption applies to both types of significance tests.

REFERENCES

Agresti, A. (1990). *Categorical data analysis,* New York: Wiley.

Bakeman, R., & Dorval, B. (1989). The distinction between sampling independence and empirical independence in sequential analysis. *Behavioral Assessment*, **11**, 31-37.

Bakeman, R., & Gottman, J. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). Cambridge: Cambridge University Press.

Bakeman, R., McArthur, D., & Quera, V. (1996). Detecting group differences in sequential association using sampled permutations: Log odds, kappa, and phi compared. *Behavior Research Methods, Instruments, & Computers*, **28**, 446-457.

Bakeman, R., & Quera, V. (1995). Log linear approaches to lag-sequential analysis when consecutive codes may and cannot repeat. *Psychological Bulletin*, **118**, 272-284.

Bakeman, R., Robinson, B., & Quera, V. (1996). Testing sequential association: Estimating exact *p* values using sampled permutations. *Psychological Methods,* **1**, 4-15.

Blair, R., & Karniski, W. (1993). An alternative method for significance testing of waveform difference potentials. *Psychophysiology,* **30**, 518-524.

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

Cervone, D. (1985). Randomization tests to determine significance levels of microanalytic congruences between self-efficacy and behavior. *Cognitive Therapy & Research*, **9**, 357-365.

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

Edgington, E. S. (1964). Randomization tests. *Journal of Psychology*, **57**, 445-449.

Edgington, E. S. (1975). Randomization tests for one-subject operant experiments. *Journal of Psychology*, **90**, 57-68.

Edgington, E. S. (1980a). Overcoming obstacles to single-subject experimentation. *Journal of Educational Statistics*, **5**, 261-267.

Edgington, E. S. (1980b). Validity of randomization tests for one-subject experiments. *Journal of Educational Statistics*, **5**, 235-251.

Edgington, E. S. (1995). *Randomization tests* (3rd ed.). New York: Dekker.

Edgington, E. S. (1996). Randomized single-subject experimental designs. *Behavior Research & Therapy*, **34**, 567-574.

Ferron, J., & Onghena, P. (1996). The power of randomization tests with responsive single-case designs. *Journal of Experimental Education*, **64**, 231-239.

Ferron, J., & Ware, W. (1994). Using randomization tests with responsive single-case designs. *Behavioral Research Therapy*, **32**, 787-791.

Good, P. (2000). *Permutation tests: A practical guide to resampling methods for testing hypotheses.* New York: Springer-Verlag.

Hayes, A. (1996). Permutation test is not distribution free: Testing the null hypothesis: $p = 0$. *Psychological Methods*, **1**, 184-198.

Knoke, D., & Burke, P. (1980). *Log-linear models* (Sage University series on quantitative applications in the social sciences, 07-020). Beverly Hills: Sage.

Kratochwill, T., Alden, K., Demuth, D., Panicucci, C., Arntson, P., McMurray, N., Hempstead, J., & Levin, J. (1974). A further consideration in the application of an analysis of variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis*, **7**, 629-633.

Krokoff, L. J., Gottman, J. M., & Roy, A. (1988). Blue collar and white collar marital interaction and communication orientation. *Journal of Social & Personal Relationships*, **5**, 201-221.

Levin, J., Marascuilo, L., & Hubert, L. (1978). $N = 1$ nonparametric randomization tests. In T. Kratochwill (Ed.), *Single subject research: Strategies for evaluating change* (pp. 167-196). New York: Academic Press.

Marascuilo, L., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences.* Monterey, CA: Brooks/Cole.

McLeod, R. S., Taylor, D. W., Cohen, A., & Cullen, J. B. (1986). Single patient randomized clinical trials: Its use in determining optimal treatment for patients with inflammation of a Kock continent ileostomy reservoir. *Lancet*, **29**, 726-728.

Onghena, P., & Van Damme, G. (1994). SCRT 1.1: Single-case randomization tests. *Behavior Research Methods, Instruments, & Computers*, **26**, 369.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, **4**, 119-130.

Tansey, R., White, M., Long, R., & Smith, M. (1996). A comparison of loglinear modeling and logistic regression in management research. *Journal of Management*, **22**, 339-358.

van den Brink, W. (1988). The robustness of the *t* test of the correlation coefficient and the need for simulation studies. *British Journal of Mathematical & Statistical Psychology*, **41**, 251-256.

Wampold, B. E., & Worsham, N. (1986). Randomization tests for multiple-baseline designs. *Behavioral Assessment*, **8**, 135-143.

Yoder, P., & Feurer, I. (2000). Quantifying the magnitude of sequential association between events and behaviors. In T. Thompson,

D. Felce, & F. Symons (Eds.), *Behavioral observation: Technology and applications in developmental disabilities* (pp. 317-334). Baltimore: Brookes.

Yoder, P. J., Klee, T., Hooshyar, N., & Schaffer, M. (1997). Correlates and antecedents of maternal expansions of utterances of children with language disabilities. *Clinical Linguistics & Phonetics*, **12**, 23-41.

Yoder, P., Short-Meyerson, K., & Tapp, J. T. (2000, April). *Sequential analysis of adult–child interactions: Issues that are rarely discussed.* Poster session presented at the Conference on Treatment Efficacy, Nashville, TN.

(Manuscript received September 8, 2000;
revision accepted for publication May 16, 2001.)