Notes and Comment

Randomization test for coupled data

BRUNO D. ZUMBO

University of Northern British Columbia,
Prince George, British Columbia, Canada

Coupled data arise in perceptual research when subjects are contributing two scores to the data pool. These two scores, it can be reasonably argued, cannot be assumed to be independent of one another; therefore, special treatment is needed when performing statistical inference. This paper shows how the Type I error rate of randomization-based inference is affected by coupled data. It is demonstrated through Monte Carlo simulation that a randomization test behaves much like its parametric counterpart except that, for the randomization test, a negative correlation results in an inflation in the Type I error rate. A new randomization test, the couplet-referenced randomization test, is developed and shown to work for sample sizes of 8 or more observations. An example is presented to demonstrate the computation and interpretation of the new randomization test.

In perceptual research, it is not unusual for researchers to report the number of organs (e.g., ears, eyes) tested, rather than the number of subjects. In these cases, subjects are then contributing two scores to the data pool. It can be reasonably argued that these two scores are not independent (i.e., uncorrelated) of one another. Data arising from such research should be referred to as *coupled* since each subject contributes a couplet of scores, and the correlation between these scores should be referred to as the *intracouple correlation*. This issue of coupled scores applies to audition, vision, and hemispheric laterality research, and to any situation in which two lateral measures are made on one subject. Therefore, a defining characteristic of coupled data is that there may be twice as many scores as there are subjects (i.e., there are n scores and $n/2$ subjects). Because the commonly used statistical inferential methods assume that the n scores are independent, a potential problem may arise when researchers base their statistical analyses on the n scores, ignoring that they arise from $n/2$ subjects. How, then, is one to perform inferential tests on data that are, potentially, highly interdependent (i.e., coupled data)?

Two quite distinct inferential approaches have been used for analyzing data arising from behavioral and social

experimentation: (1) parametric statistical methods such as t tests and analyses of variance (ANOVAs), and (2) non-parametric distribution-free methods and, particularly, randomization-based inference. Though the parametric approach is by far the most popular approach to analyzing experiments, the randomization approach is increasingly being advocated as an alternative (e.g., Edgington, 1987; Manly, 1991). The purpose of this paper is to present a new method of randomization-based inference appropriate for coupled data. Although the paper will focus on the simple two-group completely randomized design, the approach is easily generalizable to more complex experimental designs. A subtheme of the paper is that, in the case of positive intracouple correlation, the conventional randomization test performs like the parametric test—that is, the Type I error rate is inflated. In contrast, for a negative intracouple correlation, the Type I error rate of the parametric test is reduced, whereas that of the randomization test is inflated. Before describing the new randomization test, the expression *coupled data* will be further defined, and an example of coupled data will be presented. In closing, the new randomization test will be applied to the example data, and some advantages of the randomization-based inference will be discussed.

COUPLED DATA

Coupled data arise in those cases in which the observations in a study are not independent random variables but rather are pairwise related. Coupled data can be described with an example from the area of auditory perception research. Coren and Hakstian (1990) and Coren (1989) state that there appears to be a presumption among many researchers in auditory perception that the two ears are statistically independent. Hence, it is not unusual for researchers to report the number of ears tested, rather than the number of subjects. Furthermore, in the vast majority of such instances, subjects are contributing two auditory threshold profiles to the data pool, and these generally have been statistically treated as if they were not different from two ears from two different observers. The statistical problem discussed by Coren and his associates has also been noted in vision research (Ederer, 1973; Rosner, 1982) and studies of dyadic relationships (Kenny, 1988, 1995), and it could conceivably occur in laterality studies, twin studies, or any experimental or quasi-experimental setting in which the assumption of independence within groups is violated by paired or, as I will refer to them, coupled data.¹

An Example

To illustrate the issue, consider the data from a two-group completely randomized design given in Table 1. The data

The author would like to acknowledge the feedback of two anonymous reviewers, Sherry Beaumont, Bob Pruzek, David Kenny, Michael Harwell, Josef Ackerman, and, particularly, Neil Macmillan on earlier versions of this paper. Correspondence should be addressed to B. D. Zumbo, Psychology Program, University of Northern British Columbia, 3333 University Way, Prince George, BC, Canada V2N 4Z9 (e-mail: zumbob@unbc.edu).

Table 1
Example Data

Subject	Right	Left
Young-Old		
1	25	45
2	10	15
3	-5	10
4	20	70
5	5	5
6	15	25
7	35	45
Old-Old		
8	80	75
9	75	70
10	110	110
11	50	50

were kindly supplied by Professor Stanley Coren of the University of British Columbia. The data reported in Table 1 are a subset of the data reported in Coren (1989) and Coren and Hakstian (1992). The data are pure-tone hearing thresholds in the form of decibel hearing loss and range from -5 to 110 dB. A full description of the apparatus and measurement procedures used to attain these data can be found in Coren (1989) and Coren and Hakstian (1992).

In our example scenario, a gero-psychologist is interested in investigating whether there is a difference in pure-tone hearing threshold sensitivity at the center of the speech range (2000 Hz) between "young-old" (ages 70-75) and "old-old" (ages 80-85) subjects. The research question under investigation is whether the threshold hearing sensitivity (at the center of the speech range) will be greater for old-old subjects than for young-old subjects. The hearing detection threshold difference (or difference in hearing loss) between the two age groups may arise from structural and functional changes that occur in the auditory mechanisms as one ages.

Given the research question under investigation, the statistical test is of a directional hypothesis (i.e., a one-tailed test). There are 4 females and 3 males in the young-old group, and 3 females and 1 male in the old-old group. The mean age of the young-old subjects is 72.9 years (range, 70-75); the mean age of the old-old subjects is 81.8 years (range, 81-83). The Pearson correlation (i.e., the intracouple correlation) between the scores on the left and right ears is 0.901 for all of the subjects (0.736 for young-old and 0.993 for old-old). The mean threshold sensitivity is 22.86 (computed from the 14 observations) for the young-old group and 77.50 (computed from the 8 observations) for the old-old group.

The researchers are not interested in differences in the sensitivity of subjects' right and left ears; therefore, they have a total of 14 observations (i.e., 7 couplets) in the young-old group and 8 observations (i.e., 4 couplets) in the old-old group. Traditionally, in the auditory perception literature, this has been envisioned as a two-group completely randomized design and analyzed with a parametric statistical test (e.g., in this case, the independent samples *t* test

with 20 degrees of freedom) treating the data arising from the two ears as if they were independent (for examples, see Coren & Hakstian, 1990).

The problem in dealing with these coupled data in this way is that, for parametric tests, a violation of within-group independence can invalidate the statistical test (Coren & Hakstian, 1990; Kenny, 1995; Zumbo & Zimmerman, 1991). More precisely, it can be shown mathematically that, for *t* tests and ANOVAs, a positive correlation within couples results in an inflation in Type I error rate, and a negative correlation results in a reduction in Type I error rate. Therefore, if the data from the two ears are positively correlated, the Type I error rate is inflated; however, if the data from the two ears are negatively correlated, the Type I error rate is deflated.

More formally, from Zumbo and Zimmerman's (1991) Equation 2, we can derive a function showing how the Type I error rate is altered by coupled data. Let us denote α as the nominal Type I error rate of the *t* test (usually .05) and ϵ as the actual Type I error rate if we were to conduct the *t* test incorrectly ignoring the coupled data, $n = n_1 = n_2$ denotes the common sample size, and ρ denotes the intracouple correlation. We can then write the function

$$t_{\epsilon} = t_{\alpha} \sqrt{\frac{n - (1 + \rho)}{(n - 1)(1 + \rho)}}. \quad (1)$$

Three points are noteworthy from the above equation. First, the amount that the Type I error rate is altered is a function of both the magnitude of the intracouple correlation and the sample size. Second, for a fixed sample size, when $\rho = 0$, $t_{\epsilon} = t_{\alpha}$; as ρ approaches -1 in the limit, t_{ϵ} becomes larger than t_{α} ; and as ρ approaches +1 in the limit, t_{ϵ} becomes smaller than t_{α} . For example, for a nondirectional hypothesis test with 18 degrees of freedom, $t_{\alpha} = 2.10$; if $\rho = 0$, then (as expected) $t_{\epsilon} = 2.10$; for $\rho = -.99999$, $t_{\epsilon} = 683.3$; and for $\rho = .99999$, $t_{\epsilon} = 1.44$. Generally, if $\rho = 0$, then $\epsilon = \alpha$; a negative ρ would result in $\epsilon < \alpha$, and a positive ρ would result in $\epsilon > \alpha$. Finally, given that the distribution of *t* scores and the distribution of *F* scores are related by $t^2 = F$, these results generalize to the fully randomized design ANOVA, where

$$F_{\epsilon} = F_{\alpha} \left(\frac{n - (1 + \rho)}{(n - 1)(1 + \rho)} \right).$$

The mathematical details for the above discussion are omitted for lack of space, but they are available from the author (for a discussion of a more general case of within-group nonindependence, see Kenny & Judd, 1986, and Zimmerman, Williams, & Zumbo, 1993).

CONVENTIONAL RANDOMIZATION TEST

The concept of randomization-based inference was originated by Fisher (1925, section 24) and further developed mathematically by Pitman (1937a, 1937b, 1938). In short, randomization tests are a distribution-free method of hypothesis testing—that is, they make no assumptions

about population shapes. The p values that arise from randomization tests are not obtained by referring to a theoretical sampling distribution, such as the t or F distribution. Instead, the p values are obtained by reordering the data at hand (i.e., permuting the data) in a manner consistent with the experimental design.² Commonly, the set of rearrangements of the data is denoted as the referent distribution, because it is the distribution that one refers to when computing the p value.

Randomization tests are carried out with the aid of a computer in the following manner. First, a test statistic (e.g., in the two-group case, a t value or, equivalently, the difference between means) is computed for the experimental data. The data are then permuted, and the test statistic is computed for each of the resulting data permutations. The set of test statistics, including the one representing the obtained results, constitutes the referent distribution. The proportion of data permutations in the referent distribution with test statistics greater than or equal to (or for certain test statistics and hypotheses, less than or equal to) the experimentally obtained test statistic results in the p value.

In summary, these tests are often considered alternatives to parametric statistical tests. As such, Edgington (1987) states that randomization tests do not rely on the assumptions that the standard parametric statistical methods require, such as normality and random sampling.

These tests do, however, make the assumption of exchangeability. Exchangeability is seldom, if ever, discussed in behavioral science statistics textbooks or courses. One way of conceptualizing exchangeability is that, for a randomization test to be valid, every rearrangement of the data should have the same likelihood of arising. More formally, a sequence $X_1, X_2, \dots, X_n$ of variables (i.e., random variables) is said to be exchangeable if the probability of any observed sequence is unaltered when that sequence is permuted in any of the $n!$ possible ways. Then, a sufficient but not necessary condition for exchangeability is the presence of a randomization mechanism, such as random assignment. Furthermore, the exchangeability assumption allows us to use randomization tests in a context where there is no random assignment (e.g., correlational or quasi-experimental designs), such as the gero-psychology example above. The important point to note is that exchangeability may not be true with coupled data.

Very little investigation has been devoted to examining the effect of dependence within groups on the Type I error rate of the conventional randomization test. A Monte Carlo study by Harwell (1991) showed that the Type I error rate for the conventional randomization test is inflated as much as that of the t test when all of the possible pairings of observations within a group are equally correlated. This is a far more extreme case of dependence than that with coupled data; however, it suggests that the conventional randomization test may perform as poorly as the t test with coupled data. See Zumbo and Zimmerman's (1991) discussion of their Table 1 for clarification of a correlation among all of the possible pairings (which Zumbo and Zimmerman refer to as *exchangeable random variables*) versus coupled data.

Study 1: A Monte Carlo Study of the Conventional Randomization Test

Method

The Monte Carlo technique used in this investigation was designed to estimate the Type I error rates of the conventional randomization test with coupled data.

The computer-generation methodology was the same as that used in Zumbo and Zimmerman (1991). Random samples from the standard normal distribution were generated on the basis of uniform pseudorandom numbers. These normal deviates were generated using the method of Box and Muller (1958). The relation used was $X = \sqrt{-2\log_e X_1} (\cos 2\pi X_2)$, where X_1 and X_2 were independent uniform pseudorandom numbers on the interval $[0, 1]$. The random-number generator used in this investigation was the thoroughly tested prime-modulus multiplicative congruential generator described by Lewis and Orav (1989) and Lewis, Goodman, and Miller (1969). Furthermore, the adequacy of the random-number generator was evaluated using tests by Lehman (1977), Morgan (1984), and Ripley (1987). The uniform generator met the accepted requirements of rectangularity, sequential independence, and lack of patterns in a sequence of numbers.

The Type I error rate was calculated for the nominal value of .05 on data from independent populations of homogeneous variance; both directional and nondirectional hypotheses were investigated. The number of replications was 5,000. The estimated Type I error rate was the proportion of replications in which the statistical test was significant. I have chosen not to use approximate randomization tests in this study because the complete randomization distribution allowed me to discount any possible confounding effects of the approximation.³

For each replication, n_1 values were generated according to the standard normal distribution; n_2 values were generated from the same distribution. These sample values were then transformed to induce a variance-covariance pattern given in Matrix B of Zumbo and Zimmerman's (1991) Table 1. Correlations were induced by adding a multiple of one random variable to another random variable. The multiplicative constant was chosen to obtain the desired correlation coefficient, and the variables were further transformed to have mean 0 and variance 1. More specifically for positive correlation, if Z , W_1 , and W_2 are independent random variables with mean 0 and variance 1, then $X = (W_1 + bZ) / \sqrt{1 + b^2}$ and $Y = (W_2 + bZ) / \sqrt{1 + b^2}$, where $b = \sqrt{r/(1-r)}$, are dependent random variables related by $\rho(X, Y) = r$, $E(X) = E(Y) = 0$, and $\sigma_X^2 = \sigma_Y^2 = 1$. For a negative correlation, the process is exactly the same except that $Y = (W_2 - bZ) / \sqrt{1 + b^2}$, and the absolute value of r is used in the computation of b .

Performance index. Although some heuristic alternatives have been offered (e.g., Bradley, 1978), this author advocates a statistically based methodology to analyze Monte Carlo results. To take sampling error into account, an interval based on $2\sqrt{\alpha(1-\alpha)/S}$, where α is the nominal Type I error rate and S is the number of replications, was established (Harwell, 1991). Therefore, empirical Type I error rates not within the interval 0.056–0.044 reflect cases wherein the empirical Type I error is statistically different from the nominal Type I error rate.

Simulations were conducted to estimate empirically the Type I error rates of the conventional randomization test with the intracouple correlation, ρ , equal to 0, .20, .40, .60, .80, .90, and 1.0. Note, however, that due to constraints in the simulation methodology, we could not run $\rho = 1.0$ because it would result in dividing by zero when computing b . However, we were able to obtain very close approximations to $\rho = 1.0$ by using $\rho = .999999$. Sample sizes of 4, 6, 8, and 10 observations per group (i.e., 2, 3, 4, and 5 couplets per group, respectively) were used, and all tests were conducted at an alpha of .05.

Results and Conclusions

Figure 1 presents the results of the investigation into the Type I error of the conventional randomization test with

Figure 1. Empirical Type I error rates for the conventional randomization test on coupled data. Note that n s of 4, 6, 8, and 10 observations per group are equivalent to 2, 3, 4, and 5 couplets per group, respectively.

coupled data. In Figure 1, the horizontal dashed lines represent the upper and lower intervals wherein the empirical Type I error rates are not statistically different from the nominal alpha, and the vertical line traces ρ equal to zero. The intersection of the vertical line and the horizontal interval is the range of the empirical Type I error for the independent data case.

First, Figure 1 shows that, for positive correlation, the Type I error is inflated. This is similar to the effect that coupled data has on t tests and ANOVAs (Zumbo & Zimmerman, 1991). However, unlike the results with t tests and ANOVAs, a negative correlation also results in an inflation of the Type I error rate—in fact, the lines are somewhat symmetrical about $\rho = 0$. The t -test results in Figure 1 were obtained using Equation 1 with 18 degrees of freedom. This finding is consistent for both directional and

nondirectional hypotheses. It appears, then, that (as suggested by Harwell's, 1991, findings) the conventional randomization test should not be used with coupled data.

COUPLET-REFERENCED RANDOMIZATION TEST

The findings in Study 1 and those in Harwell (1991) suggest that a new randomization test is needed for coupled data. This new randomization test is developed from the following principle: Rather than forming permutations at the level of the individual observation (i.e., conventional randomization test), if one were to form the data rearrangements on the basis of the coupled data structure, then they will be exchangeable. That is, because the experimental unit that is being randomized is the couplet, one must per-

Table 2
Example of Conventional and
Couplet-Referenced Randomization Tests

Permutation	Group 1	Group 2
Conventional Randomization Test		
1	X_1, X_2, X_3, X_4	X_5, X_6, X_7, X_8
2	X_1, X_2, X_3, X_5	X_4, X_6, X_7, X_8
3	X_1, X_2, X_3, X_6	X_4, X_5, X_7, X_8
4	X_1, X_2, X_3, X_7	X_4, X_5, X_6, X_8
5	X_1, X_2, X_3, X_8	X_4, X_5, X_6, X_7
6	X_1, X_2, X_5, X_4	X_3, X_6, X_7, X_8
7	X_1, X_2, X_6, X_4	X_3, X_5, X_7, X_8
.	.	.
.	.	.
70	X_5, X_6, X_7, X_8	X_1, X_2, X_3, X_4
Couplet-Referenced Randomization Test		
1	X_1, X_2, X_3, X_4	X_5, X_6, X_7, X_8
2	X_1, X_2, X_5, X_6	X_3, X_4, X_7, X_8
3	X_1, X_2, X_7, X_8	X_3, X_4, X_5, X_6
4	X_3, X_4, X_5, X_6	X_1, X_2, X_7, X_8
5	X_3, X_4, X_7, X_8	X_1, X_2, X_5, X_6
6	X_5, X_6, X_7, X_8	X_1, X_2, X_3, X_4

mute the observations to obtain a couplet-referenced referent distribution in order for the randomization test to be valid. Hence, this new test statistic is referred to as a *couplet-referenced randomization test*.

The couplet-referenced randomization test will be demonstrated with the aid of Table 2. The data for Table 2 represent the observations from a two-group completely randomized design involving 4 subjects, wherein X_1 and X_2 are from Subject 1, X_3 and X_4 are from Subject 2, X_5 and X_6 are from Subject 3, and X_7 and X_8 are from Subject 4. The upper part of Table 2 shows the referent distribution for the conventional randomization test. The 8 observations are permuted on the basis of group membership. The resulting randomization test would involve, for example, 70 rearrangements of the data with 4 observations per group. The bottom part of Table 2 shows the referent distribution for the couplet-referenced randomization test. The eight observations are permuted on the basis of group membership; however, the observations from each subject are not separated. The resulting randomization test would involve 6 rearrangements with 4 observations per group. A computationally convenient way to obtain the couplet-referenced randomization test is to average across the scores for an individual and then perform a conventional randomization test on those averages. The resulting statistical decision is formally equivalent to the result of the couplet-referenced randomization test. The details of this equivalence are omitted for lack of space, but they are available from the author.

Study 2: A Monte Carlo Study of the Couplet-Referenced Randomization Test

Method

The Monte Carlo technique used in this investigation was designed to estimate empirically the Type I error rates of the couplet-referenced

randomization test with coupled data. The computer-generation methodology, performance index, and sample sizes were the same as those in Study 1.

Results and Conclusions

Figure 2 presents the results of the investigation into the Type I error of the couplet-referenced randomization tests with coupled data. Again, the horizontal dashed lines represent the upper and lower intervals wherein the empirical Type I error rates are not statistically different from the nominal Type I error rate, and the intersection of the vertical line and the horizontal interval is the range of the empirical Type I error for the independent case.

Examining Figure 2, it is apparent that the couplet-referenced randomization test generally performs well for sample sizes as small as 10 observations (5 couplets) per group for both directional and nondirectional hypotheses. However, the test is conservative for sample sizes of 8 observations (4 couplets) per group, but less so for directional hypotheses. Moreover, the couplet-referenced randomization test performs dismally (an empirical Type I error rate of zero) for sample sizes of 4 and 6 observations (i.e., 2 and 3 couplets, respectively) per group. This poor performance can be accounted for by the fact that the referent distribution of the randomization test is discrete (Lehmann, 1975). Theoretically, by using extraneous randomization, we could correct this problem and conduct the statistical test at any significance level, but this is not a commonly used technique in practice. Therefore, for the common Type I error rate of .05, the couplet-referenced randomization test is useful for sample sizes of 8 observations (i.e., 4 couplets) and larger per group. Note that, for unequal sample sizes, these guidelines apply for the larger sample size.

ANALYSIS OF THE EXAMPLE DATA

With the results of the two simulation studies in mind, we will now analyze the gero-psychology data. As a preliminary examination of the data, a Dallal plot (Berk, 1994, p. 162) was created and is listed in Figure 3. A Dallal plot is a box plot superimposed over a density plot. The circles show the location of each observation, and the box plots show the quartiles for each group. Given that the box plot for the old-old group is much higher than the box plot for the young-old group, it is suggestive that there is a significant difference between the means of the two groups. Furthermore, the combined information from the box and density plots shows that the young-old group is positively skewed and that it contains what appears to be 1 aberrant observation. The drastically non-normal shape of the distribution suggests that the modified t test (a parametric test) introduced by Zumbo and Zimmerman (1991) may be inappropriate and that a nonparametric/distribution-free test such as the randomization test may be more appropriate (Zimmerman & Zumbo, 1993). Therefore, the data from Table 1 were submitted to a couplet-referenced randomization test to test the hypothesis $H_0: \mu_1 - \mu_2 \geq 0$ versus

Figure 2. Empirical Type I error rates for the couplet-referenced randomization test. Note that n s of 4, 6, 8, and 10 observations per group are equivalent to 2, 3, 4, and 5 couplets per group, respectively.

$H_a: \mu_1 - \mu_2 < 0$, as described in Table 2, with a resulting sample mean difference of 54.64 and a one-tailed exact probability of .003. This was not an approximate randomization test but rather was based on all 330 permutations and was computed with the program PITMAN (Dallal, 1988). The conclusion, then, is that the threshold sensitivity of the old-old group was statistically significantly greater than that of the young-old group. This finding also has substantive meaning in that the mean threshold sensitivity for the young-old group can be categorized as "mild to moderate hearing loss" as opposed to the mean of the old-old group, which can be categorized as "severe to profound hearing loss." It is important to note that the above hypothesis test is distribution-free, so the apparent non-

normality seen in Figure 3 is not as much of a concern with this test as it would be with parametric tests. Most importantly, when analyzing coupled data, the probability of Type I error is not inflated above its nominal level (.05) with the couplet-referenced randomization test.

SOME CONCLUDING REMARKS

The purpose of this paper was to present a new randomization test as well as some Monte Carlo simulations with the couplet-referenced and conventional randomization tests on coupled data. The findings of the Monte Carlo simulations can be summarized in two parts. First, the results regarding the conventional randomization test

Figure 3. Dallal plot of the example data. For the age groups, 1 = young-old, and 2 = old-old.

indicate that using it on coupled data leads to an inflated number of false rejections of the null hypothesis, well over the nominal Type I error probability (and much the same as the parametric tests on coupled data). However, for negative correlations, parametric tests result in a deflation of Type I error rate, whereas the conventional randomization test results in an inflation of Type I error rate. It should be noted that a statistical test with a Type I error rate less than its nominal value is still valid (Lehmann, 1975), but one needs to be concerned with a possible accompanying reduction in statistical power. Therefore, the t test is, strictly speaking, valid for negative correlation, whereas the conventional randomization test is not. Second, it was shown that the couplet-referenced randomization test performs adequately for a sample size as small as 8 or 10 observations (i.e., 4 or 5 couplets) per group.

The couplet-referenced randomization test is an example of the true advantage of randomization tests over the more commonly used parametric statistical tests (e.g., t and F tests). That is, the true potential of randomization tests is not their potential in supplanting currently available tests but rather their use in developing new statistical tests for problems that, for many reasons, appear mathematically complex and/or intractable. For example, it was stated earlier that coupled-data scenarios include dyadic studies, or any experimental or quasi-experimental settings in which the assumption of independence within groups is violated by coupled data. The parametric test presented by Zumbo and Zimmerman (1991) requires an estimate of the correlation among the observations. This correlation is relatively easy to calculate for lateral measures, such as for ears or eyes, that can be clearly categorized as left or right. However, if one is in a research context where elements of the couplets cannot be so easily categorized (e.g., familial, twin, dyadic relationships, or paired interview data), then an intraclass correlation needs to be estimated and used in the Zumbo and Zimmerman formula. Estimation of the intraclass correlation, however, is computationally de-

manding, and various competing estimation methods are available (Srivastava, Keen, & Katapa, 1988). Furthermore, the Zumbo and Zimmerman formula was derived assuming a population value for the correlation coefficient; therefore, a sample estimate makes the Zumbo and Zimmerman formula an approximate test (see Zumbo & Zimmerman, 1991, for a discussion and some results regarding the effect of this approximation). Therefore, another advantage of the couplet-referenced randomization test is that one does not have to estimate the intraclass correlation. Finally, these randomization tests are distribution-free, so they apply to, as Pitman notes, any population distribution.

REFERENCES

- BERK, K. N. (1994). *Data analysis with Student SYSTAT*. Cambridge, MA: Course Technology.
- BOX, G. E. P., & MULLER, M. (1958). A note on the generation of normal deviates. *Annals of Mathematical Statistics*, **29**, 610-611.
- BRADLEY, J. V. (1978). Robustness? *British Journal of Mathematical & Statistical Psychology*, **31**, 144-152.
- COREN, S. (1989). Summarizing pure-tone hearing thresholds: The equipollence of components of the audiogram. *Bulletin of the Psychonomic Society*, **27**, 42-44.
- COREN, S., & HAKSTIAN, A. R. (1990). Methodological implications of interaural correlation: Count heads not ears. *Perception & Psychophysics*, **48**, 291-294.
- COREN, S., & HAKSTIAN, A. R. (1992). The development and cross-validation of a self-report inventory to assess pure-tone threshold hearing sensitivity. *Journal of Speech & Hearing Research*, **35**, 921-928.
- DALLAL, G. E. (1988). PITMAN: A FORTRAN program for exact randomization tests. *Computers & Biomedical Research*, **21**, 9-15.
- DWASS, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, **28**, 181-187.
- EDERER, F. (1973). Shall we count number of eyes or number of subjects? *Archives of Ophthalmology*, **89**, 1-2.
- EDGINGTON, E. S. (1987). *Randomization tests* (2nd ed.). New York: Marcel Dekker.
- FISHER, R. A. (1925). *Statistical methods for research workers* (1st ed.). Edinburgh: Oliver & Boyd.
- HARWELL, M. R. (1991). Using randomization tests when errors are unequally correlated. *Computational Statistics & Data Analysis*, **11**, 75-85.
- KELLER-MCNULTY, S., & HIGGINS, J. J. (1987). Effect of tail weight and outliers on power and Type-I error of robust permutation tests of location. *Communications in Statistics: Part B. Simulation & Computation*, **16**, 17-35.
- KEMPTHORNE, O. (1955). The randomization theory of experimental inference. *Journal of the American Statistical Association*, **50**, 946-967.
- KENNY, D. A. (1988). The analysis of data from two-person relationships. In S. Duck (Ed.), *Handbook of interpersonal relations* (pp. 57-77). London: Wiley.
- KENNY, D. A. (1995). The effect of nonindependence on significance testing in dyadic research. *Personal Relationships*, **2**, 67-75.
- KENNY, D. A., & JUDD, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin*, **99**, 422-431.
- LEHMAN, R. S. (1977). *Computer simulation and modeling: An introduction*. Hillsdale, NJ: Erlbaum.
- LEHMANN, E. L. (1975). *Nonparametrics*. San Francisco: Holden-Day.
- LEWIS, P. A. W., GOODMAN, A. S., & MILLER, J. M. (1969). A pseudo-random number generator for the System 360. *IBM Systems Journal*, **8**, 136-146.
- LEWIS, P. A. W., & ORAV, E. J. (1989). *Simulation methodology for statisticians, operations analysts, and engineers* (Vol. 1). Pacific Grove, CA: Wadsworth.
- MANLY, B. F. J. (1991). *Randomization and Monte Carlo methods in biology*. New York: Chapman & Hall.

- MORGAN, B. J. T. (1984). *Elements of simulation*. London: Chapman & Hall.
- PITMAN, E. J. G. (1937a). Significance tests which may be applied to samples from any population. *Journal of the Royal Statistical Society*, **3**, 119-130.
- PITMAN, E. J. G. (1937b). Significance tests which may be applied to samples from any population II. *Journal of the Royal Statistical Society*, **3**, 225-232.
- PITMAN, E. J. G. (1938). Significance tests which may be applied to samples from any population III: The analysis of variance test. *Biometrika*, **29**, 322-335.
- RIPLEY, B. D. (1987). *Stochastic simulation*. New York: Wiley.
- ROSNER, B. (1982). Statistical methods in ophthalmology: An adjustment for the intraclass correlation between eyes. *Biometrics*, **38**, 105-114.
- SRIVASTAVA, M. S., KEEN, K. J., & KATAPA, R. S. (1988). Estimation of interclass and intraclass correlations in multivariate familial data. *Biometrics*, **44**, 141-150.
- ZIMMERMAN, D. W., WILLIAMS, R. H., & ZUMBO, B. D. (1993). Effect of nonindependence of sample observations on some parametric and nonparametric statistical tests. *Communications in Statistics: Part B. Simulation & Computation*, **22**, 779-789.
- ZIMMERMAN, D. W., & ZUMBO, B. D. (1993). Relative power of parametric and nonparametric statistical methods. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Vol. 1. Methodological issues* (pp. 481-517). Hillsdale, NJ: Erlbaum.
- ZUMBO, B. D., & ZIMMERMAN, D. W. (1991). Further evidence for Coren and Hakstian's "Methodological implications of interaural correlation: Count heads not ears" and an alternative correction formula. *Perception & Psychophysics*, **50**, 297-301.
- ZUMBO, B. D., & ZIMMERMAN, D. W. (1993). Postscript: Some closing comments on "Alternatives to classical statistics" (normal theory). *Canadian Psychology*, **34**, 440-445.

NOTES

1. The term *paired data* is avoided due to its possible confusion with the within-subject design.
2. Technically, a randomization test is a permutation test that has the randomization mechanism as the basis of its permutation (Kempthorne, 1955). Therefore, it is a permutation test whose conditional empirical sampling distribution is obtained by a complete reordering (i.e., permutations) of the observations. The referent distribution for a permutation test is conditional in the sense that it is technically valid only for the data at hand—that is, it is not necessary to make reference to a population.
3. Approximate permutation tests are those based on a less than complete reordering, usually a subsampling. In brief, as early as Dwass (1957), it has been known that there is some loss in power and accuracy resulting from not using the full number of permutations. Keller-McNulty and Higgins (1987) and Zumbo and Zimmerman (1993) acknowledge that the effect of this approximation should not be ignored and that the very important mathematical theory of how much loss in power and accuracy is expected under general conditions is not yet known. However, the results in Keller-McNulty and Higgins are promising.

(Manuscript received May 11, 1993;
revision accepted for publication August 4, 1995.)