Behavior and Social Issues, 12, 53-63 (2002). © Behaviorists for Social Responsibility

SPECIAL SECTION: STATISTICAL ANALYSES FOR SMALL SAMPLES

SMALL GROUP STATISTICS: A MONTE CARLO COMPARISON OF PARAMETRIC AND RANDOMIZATION TESTS

Chris Ninness, Richard Newton, Jamie Saxon, Robin Rumph,
Anna Bradfield, Carol Harrison, and Eleazar Vasquez III[1]
Stephen F. Austin State University

Sharon K. Ninness
Angelina College

ABSTRACT: Under some conditions, behavioral research includes statistical tests to assess the probability of experimental outcomes. Occasionally, sample sizes are insufficient to fulfill the assumptions of traditional parametric procedures. The following study assessed the statistical advantages, correspondence, and accuracy of univariate analyses when small-*n* samples are calculated in both randomized and traditional/parametric formats. Specifically, we compared probability values for both parametric and randomized two-sample independent *t*-tests across a series of Monte Carlo, pseudorandom two-group data sets of equal size. In doing so, we identified some patterns of probability change associated with decreasing group size.

Occasionally, applied and basic research in behavior analysis entails statistical tests of group data (e.g., Hayes, Bissett, Korn, Zettle, Rosenfarb, Cooper, & Grundt, 1999). However, many behavioral preparations that compare group data may not have sample sizes large enough to conform to the assumptions of normal curve theory (e.g., Miller, & Armus,1999). While behavioral research always will be based in single-subject design (Sidman, 1960), it has become increasingly common for researchers to employ small groups in their experimental preparations, particularly in studies relating to equivalence and human computer-interaction effects. For example, if researchers were interested in the effects of prior learning on equivalence class formation, outcomes might be analyzed in terms of individual correct responses and differences in group performance on various tasks (e.g., Peoples, Tierney, Bracken, & McKay, 1998). Researchers assessing the potentially reinforcing effects of various forms of computer feedback might assess behavior change in terms of individual and small group performance across conditions (e.g., Ninness, Rumph, McCuller, Bradfield, Saxon, & Calliou, 2001).

Until very recently, behavior analysts had a rather limited range of nonparametric options when employing small-*n* experimental designs. Statisticians

[1] This investigation was supported in part by a Texas Excellence Funds Grant. We gratefully acknowledge the software editing and comments provided by Michael Coffee, Morris Lang, and Jeremy Cobb in the development of this study. Correspondence concerning this article should be addressed to Chris Ninness, Department of Human Services, School & Behavioral Psychology Program, PO Box 13019 SFA Station, Stephen F. Austin State University, Nacogdoches, TX 75962 (e-mail: cninness@titan.sfasu.edu).

are the first to point out that most conventional nonparametric tests show weak sensitivity to treatment effects with small-*n* data (e.g., Todman & Dugard, 2001). Nevertheless, parametric procedures are not considered a viable option since there are limits as to how small group sizes can be before the probabilities based on parametric tests become unreliable. In fact, the authors of many basic and advanced texts admonish readers that good applied and basic research requires "fairly" large *n*'s. For example, Sowell (2001) notes that "samples of at least 30 or more participants" (p. 131) are required to have confidence in obtained statistical probabilities. Roscoe (1975) affirms that *n*'s of 20 to 25 appear to be the absolute minimum for a reasonable likelihood of finding significant differences in treatment effects. And, Hopkins, Hopkins, and Glass (1996) indicate that when sample sizes are small, the sampling error is greatly increased; it is not until sample size approaches 25 that the *t*-distribution begins to mimic the normal curve. This perspective has often inhibited the development, credibility, and publication of empirical research that employs relatively small numbers of participants (see Ninness, Rumph, and Bradfield, 2001, for a discussion).

However, there is a body of sophisticated, reliable, and precise statistical strategies that do not entail any underlying assumptions regarding random and independent sampling, homogeneity of variance, or normality. As described by Edgington (1995), a *randomization test* is a permutation test that precludes all assumptions regarding how data are collected. Randomization tests are a series of mathematical operations providing a test statistic to be repeatedly computed for all possible permutations of a given data set (Ninness, McCuller, & Ozenne, 2000). The proportion of outcomes that occur with as large a probability as those that are actually obtained determines the statistical likelihood of group differences, or trend changes, occurring by chance. Using these conservative but mathematically precise systems, researchers who calculate the probability of their research findings on a small number of subjects need not speculate as to whether their data base is of a sufficient group size or whether it has been randomly obtained. The randomization test calculates all possible ways that the results could have occurred, given the data at hand, and it identifies how likely it is that the obtained differences could have occurred by chance (see Good, 1994, for a discussion). Randomization tests, even more than parametric tests, emphasize the importance of random assignment of participants to treatments (Efron & Tibshirani, 1993). Sample size is completely irrelevant to the internal validity of the test statistic; however, as with all other statistical procedures, external validity can only be gauged by addressing the *logical probability* that other populations share the relevant characteristics of the sample.

As a practical matter, consider a study in which 8 students are selected from a special education classroom and *randomly assigned* to either direct instruction or control conditions. Following treatment, the four students receiving direct instruction obtain reading scores of 86, 90, 92, and 94, and the four controls score 70, 80, 80, and 83 respectively. If this data were run according to conventional parametric procedures, the two-tailed t-test would reject the null hypothesis at the 0.01 level of significance. While this calculation would be statistically significant, it would also be considered highly inappropriate given the small sample size. Yet, it is equally inappropriate to suppose that there is no reasonable strategy to determine the probability of the given scores occurring by chance. By precluding any assumptions that the data is representative of some larger population from which it is drawn, a randomization test will calculate that the exact probability of the above outcomes. Specifically, a randomization test will demonstrate that the exact probability of this outcome occurring by chance is 0.028 (see Edgington, 1995 for a similar example). Such an outcome makes no inferences or speculations about other populations not included in the study. The probability of these scores occurring by chance is identified exactly, and the scientific and social implications are clear for behavior analysts who conduct small-*n* studies. Unless small group researchers are able to employ powerful and precise data analyses, their findings will not have credibility in larger academic arena. Randomization tests have mathematical credibility, and they are powerful. However, until quite recently, they have not been available.

These procedures have been understood and appreciated by generations of mathematical statisticians; however, prior to the relatively recent advances in and access to high speed computers, these procedures were far too time consuming, inefficient, and expensive for most practical applications. Edgington (1995) points out that as few as 30 subjects placed into 3 groups may generate over 5 trillion data permutations that must be calculated to obtain exact levels of significance for some versions of the randomization test. Only a few years ago, this would have been an exceedingly long and expensive calculation (see Manly, 1997, for a detailed discussion). Ironically, it is still popularly believed that randomization tests entail long and unmanageable computer computations for the purposes of most applied researchers (e.g., Williams, Zumbo, & Zimmerman, 2001).

Irrespective of computational complexity and duration, statisticians have been exploring this permutation and randomization test for some time. Pitman (1937) provided much of the theoretical foundation for randomization tests and experimental permutation tests; however, he noted that the primary concepts in randomization are embedded in Fisher’s writings of the same period (Fisher, 1936; Savage, 1976). Still, only terse and indirect reference is made to these procedures in the early decades of statistical computing (e.g., Winer, 1971).

With recent improvements in processing speed, Manly (1997) has shown that randomization tests and conventional parametric procedures provide “roughly” the same probabilities if the samples approximate normality and the assumptions of homogeneity and independence are not violated. He advances research by Romano (1989) showing that randomization tests arrive at essentially the same levels of significance when large samples are employed and violations of assumptions have not occurred. Moreover, as noted by Edgington (1995), randomization tests usually generate more power than conventional procedures when data points are not normally distributed. This point has been supported by Peres-Neto and Olden (2001), who demonstrated that randomization tests produced the best type I error rates when compared to probabilities generated by classical ANOVA and the Kruskal-Wallis test.

To more fully address this issue, we developed a *Monte Carlo method* study based on uniform data. *Monte Carlo method* refers to a general type of computer simulation of behavioral or physical events for a very large number of occurrences. It is not unusual for Monte Carlo methods to run 10,000 variations on the behavior of a single variable (see Manly, 1997, for a discussion). This increasingly popular experimental tool allows researchers to model and explore more complex systems than could ever be revealed in natural contexts. Using a blend of random number generators and probability theory, behavior patterns are generated and analyzed in every conceivable (and inconceivable) distribution; however, the details of any particular Monte Carlo software system will be tied to the particular interests and needs of a given researcher. Generally, there are several commonly employed strategies in comparing statistical procedures with Monte Carlo methods. One strategy is to construct asymmetrical distributions that reflect values consistent with preestablished differences between groups. If a particular statistical test demonstrates good performance in a large number of simulations, it is reasonable to assume that the procedure would be functionally accurate and powerful when applied to real data.

Recent Monte Carlo research seems to suggest that randomization tests appear to be generally more robust to violation of the statistical assumptions associated with parametric and classic nonparametric approaches (Peres-Neto & Olden, 2001). Another strategy is to simulate and manipulate data based on previous studies that have demonstrated significance and high levels of internal validity (Taylor & Gerrodette, 1993). Alternatively, one can assume that, given a large number of computer-generated random scores, probabilities (P-values) for group differences should be normally distributed and 5% of these P-values should fall at or below a α set at .05. This system lends itself to directly correlating all outcomes obtained from different types of statistical tests. It allows direct point-by-point comparisons of all P-values obtained from two or more statistical tests, and it provides a measure of accuracy of the statistical tests in approaching specified α levels.

Presently, it is unclear as to how parametric and randomization procedures systematically correspond or differ across decreasingly small group sizes. Thus, researchers who employ small *n*'s have no way of knowing how levels of significance obtained by randomization tests might correspond with those obtained by traditional statistics. The following study assessed the statistical advantages, correspondence in probabilities, and accuracy of applied univariate statistical procedures when calculated according to both randomized and traditional/parametric formats. Specifically, we compared P-values for Randomized *t*-tests and classical Student's *t*-tests across an array of Monte Carlo generated groups.

METHOD

Apparatus and Software

The research was conducted on a Dell Inspiron 5000e laptop computer (700 MHZ processor and 512 MB RAM) with a 14-inch screen. Statistical and Monte Carlo method software was written by Chris Ninness in C++ and Visual Basic 6 for IBM PC compatible machines; however, the original *randomization test* algorithms were developed in Fortran by Edgington (1995) and adapted to run in the C++ language as developed and described by Ninness, Rumph, & Bradfield (2001). Since we employed multiple trials for each data set, our random number generator was seeded by the computer's clock, and each file in the Monte Carlo series was identified and seeded by increments to the nearest nanosecond. Algorithms were audited and refined by LCSDG, LLC. (Several of our current C++ versions of randomization test procedures are freely available to all academic researchers on-line at www.lcsdg.com/psychStats.)

Experimental Design and Procedures

This study obtained changing P-values associated with group differences using equal *n*'s of decreasing sample size. To evaluate the effect of sample size on *t*-test P-values, our pseudorandom number generator produced 1,000 sets of scores (ranging from 00 to 99) for groups with equal *n*'s of 12, 10, 8, 6, and 4. Following the development of these Monte Carlo data sets, P-values for the randomized *t*-tests and parametric *t*-tests were obtained, correlated, and graphed. The percentage of randomized *t*-test P-values and parametric *t*-test P-values that fell at or below α (set at .05) was identified for each of these distributions. To determine reliability of these algorithms, a second set of scores was generated, and the same calculations were conducted on all 5 group sizes.

RESULTS

Figure 1 shows the scatterplot, correlation, regression line, and percentages of P-values that fell at or below .05 for group sizes of 4 in each group on parametric and randomized *t*-test procedures. In the first generation of 1,000 data sets of size 4, P-values for randomized *t*-tests and parametric *t*-tests correlated at .9853. Here, 6.2% of the obtained parametric probabilities fell at or below the truncated .05 α, while only 4.3% of the randomized P-values were at or below this level. Figure 1 also shows the same information for *n*'s of 6 with a conspicuous improvement in the amount of scatter along the regression line. P-values correlated at .9987, with 5.5% of the parametric outcomes were at or below the designated α level. The randomized *t*-test obtained 4.7% of the corresponding probabilities below α. In both conditions, the randomized *t*-test probabilities that fell at or below .05 always corresponded with those identified by the parametric *t*-test; however, some of the parametric probabilities slightly exceeded the .05 (see circled areas on Figure 1).

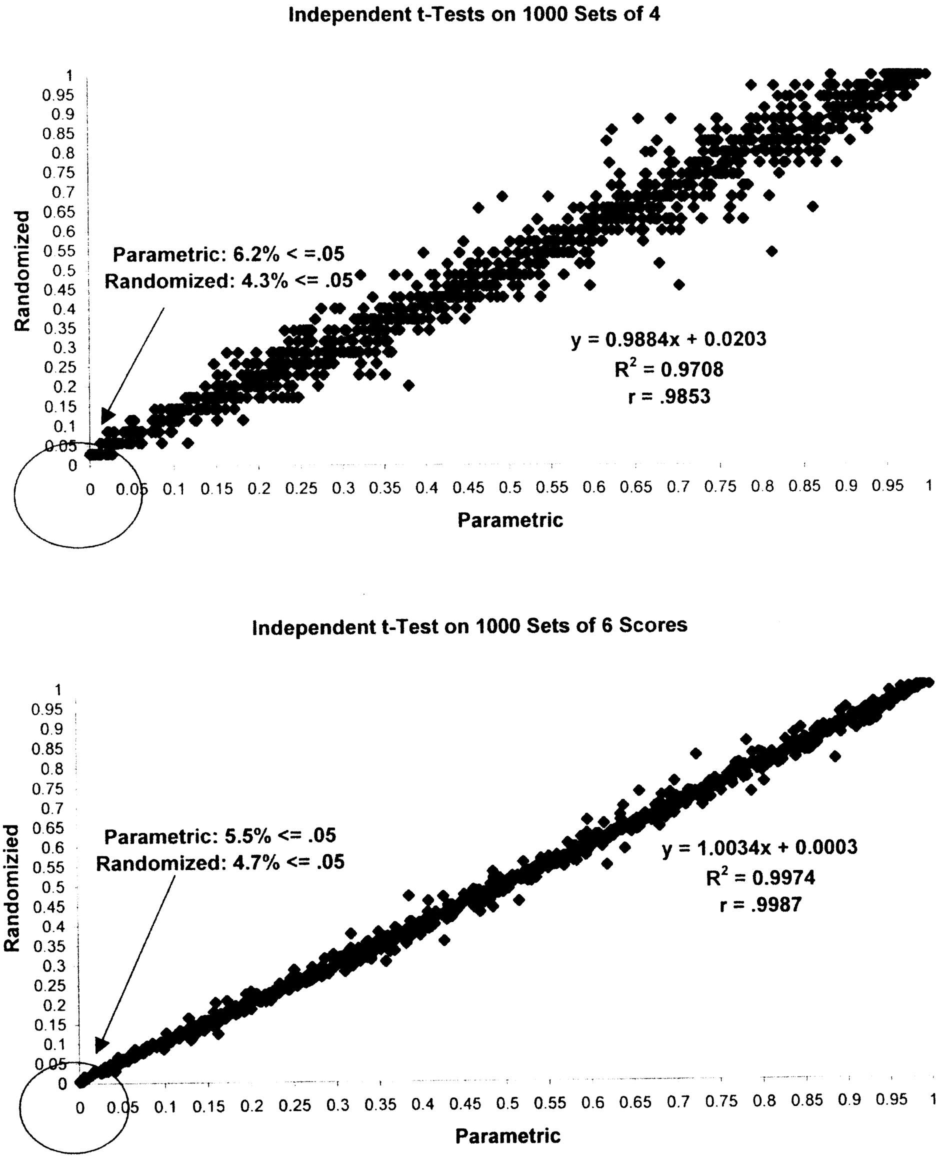

Figure 1. Monte Carlo generated randomized *t*-test and parametric *t*-test probabilities. P-values are based on 1,000 sets of 4 scores per group and 1,000 sets of 6 scores per group.

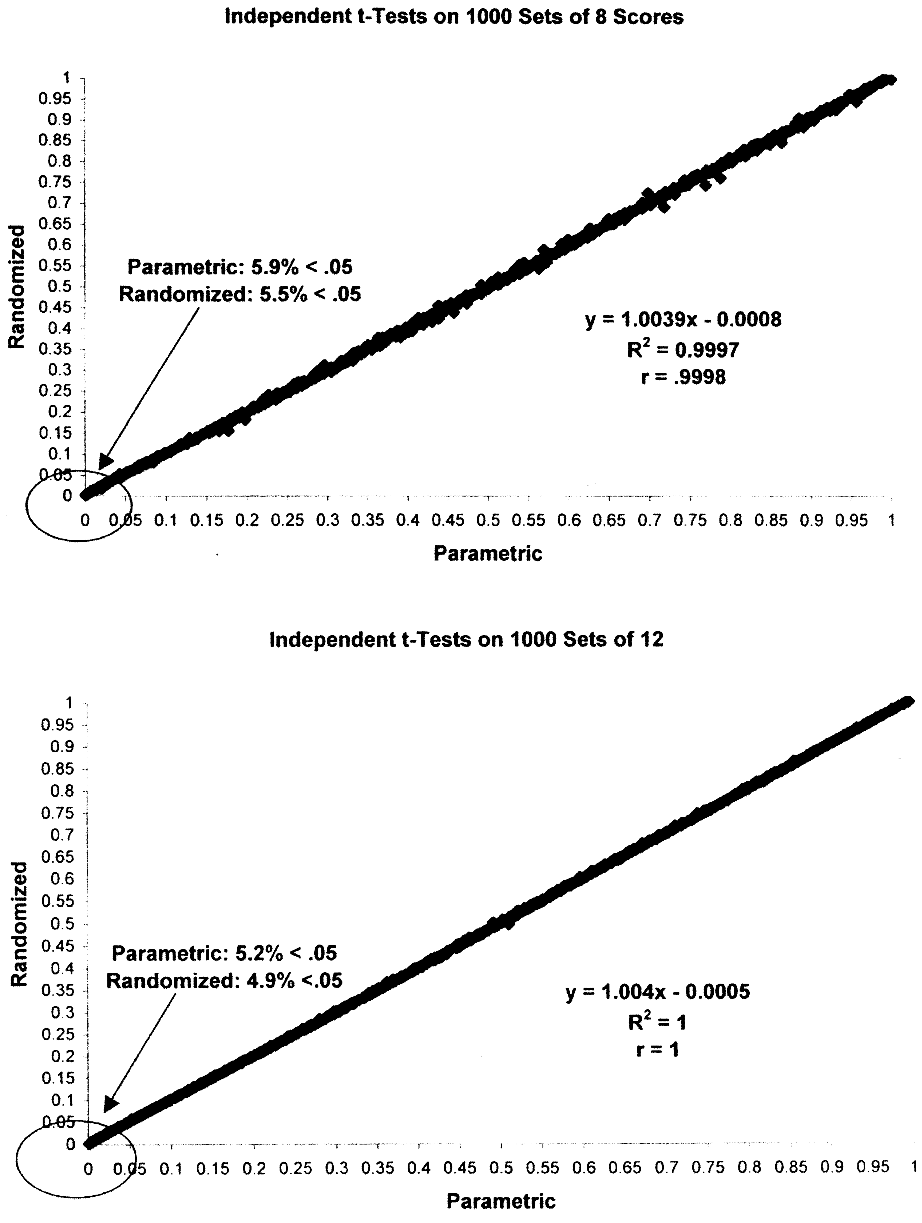

Figure 2. Monte Carlo generated randomized *t*-test and parametric *t*-test probabilities. P-values are based on 1,000 sets of 8 scores per group and 1,000 sets of 12 scores per group.

Figure 2 shows the scatterplot, correlation, regression line, and percentages of P-values that fell at or below .05 for group sizes of 8 on parametric and randomized *t*-test procedures. P-values were correlated at .9998, with 5.9% of the obtained parametric probabilities falling at or below the .05 α level and 5.5% of the randomized P-values at or below this level. Since the scatterplot produced from group sizes of 10 was not visually distinct from that of 8, it is not displayed; the correlation between parametric and randomized *t*-test probabilities was found to be .9994, and the associated α levels were found to be at 5.1% and 5%, respectively.

Figure 2, which provides information for group sizes of 12, shows no measurable scatter along the regression line. Here, P-values correlated at 1.0 (computer calculations were automatically rounded up from .9999). In this graph, 5.2% of the parametric *t*-test outcomes were at or below the designated α level. The randomized *t*-test found 4.9% of the corresponding probabilities at or below α. We replicated all of the above functions and found the resulting correlations to be within .001 of our initial calculations for all 5 group sizes.

DISCUSSION

For large-*n* data, there is very little enthusiasm in the mainstream behavioral sciences for finding alternatives to the conventional statistical tests (Todman & Dugard, 2001). However, as a researcher's sample size shrinks, so do the researcher's probabilities of finding acceptable data analysis systems, external funding and outlets for publication. The theoretical and social implications are clear to those who have tried to obtain support and disseminate their findings in the mainstream. Until quite recently, it has been gratuitously assumed that small-*n* studies could not stand up to a test of probability and a substantial portion of behavioral research has been disenfranchised by the mainstream of social sciences. High-speed computers and randomization tests have made this assumption untenable and obsolete.

Consistent with the recent outcomes of Peres-Neto and Olden (2001), we found little or no difference between randomized and parametrically derived probabilities with some group sizes. In fact, the Monte Carlo *n* of 12 produced a functionally perfect point-by-point correspondence between these approaches. However, as group sizes decreased in steps of 2, P-values fell in minor disagreement. This difference did not become conspicuous until the *n*'s were reduced to 6 and 4. Even with these very small *n's*, the correlations were at .9987 and .9853, respectively. Interestingly, the randomization P-values always agreed with those generated by the parametric Student's *t*-test, but the inverse was not true in this study. With smaller *n*'s, the parametric P-values became inflated above the designated α level, while the randomized *t*-test became more and more conservative. Thus, with *n*'s less than 12, it appears that the randomized *t*-test consistently represents a very conservative but reliable measure of type I error rates.

Results from this study demonstrate the reliable and strong correspondence between classical and randomized statistical strategies—at least in terms of the commonly employed two-sample independent *t*-test with equal *n*'s. Similar to the way in which means from increasing sample sizes from asymmetrical populations gradually assume a perfectly normal distribution of means (Hopkins et al., 1996), increasing sample sizes for randomized and parametric *t*-tests appears to gradually generate a perfect correlation between these two approaches.

While a series of strong agreements between parametric and randomized *t*-test P-values do not unequivocally prove the power of either test, they do provide a measure of "inter-test" reliability. Moreover, the fact that we repeatedly identified randomized *t*-test P-values that slightly underestimated or very closely approximated .05 α's 5% of the time, suggests that the randomized *t*-test provides a practical level of statistical precision for small group comparisons in applied and basic research.

When comparing parametric and randomized *t*-tests, sample sizes at or above 12 appear to reach an asymptotic correlation of 1 when employing algorithms using *systematic permutations*. This type of randomization test calculates levels of significance based on *all* of the possible ways in which data can be permuted within a given set of scores. The number of data permutations and amount of computer time for sample sizes above 12 quickly becomes astronomical. However, Edgington (1995) describes another, less precise *random permutation t*-test algorithm that generates P-values by randomly selecting from the set of all possible test statistics (see Manly, 1997, and Todman & Dugard, 2001, for a discussion of these algorithms). Future research might explore these algorithms in terms of their correspondence with conventional statistical tests.

The present study only addresses simulated outcomes generated with one dependent variable and two groups. Outcomes derived from the two-sample independent *t*-test may be very different from those obtained with multiple groups and multiple independent variables. Nevertheless, these data provide substantial evidence that applied and basic researchers are well positioned to reliably calculate levels of significance based on small *n*'s when using randomization tests. Nevertheless, it is important to reiterate a point emphasized by Edgington (1995): the external validity of any experiment can only be gauged by judging the logical probability that other populations share the relevant characteristics of the individuals who participated in a given study.

Given the common objection that small-*n* studies lack internal validity due to their inability to fulfill the critical assumptions of traditional sampling techniques, the various types of randomization tests we are placing on www.lcsdg.com/psychStats may be a logical alternative for behavior analysts who are interested in determining the precise probability of outcomes based on small group designs. In closing, we feel compelled to point out that none of the on-line software we are developing could have been conceivable without the brilliant algorithms originally developed by Edgington and his colleagues.

REFERENCES

Edgington, E. S. (1995). *Randomization tests.* New York: Marcel Deckker, Inc.

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Fisher, R. A. (1936). The coefficient of racial likeness and the future of craniometry. *Journal of the Royal Anthropological Institute of Great Britain and Ireland, 66,* 57-63.

Glass, G. V., Peckham, P. D. & J. R. Sanders. (1972). Consequences of failure to meet assumptions underlying the fixed-effects analysis of variance and covariance. *Review of Educational Research, 42*, 237-288.

Good, P. (1994). *Permutation tests: A practical guide to resampliing methods for testing hypotheses.* New York: Springer-Verlag.

Hayes, S. C., Bissett, R. T., Korn, Z., Zettle, R. D., Rosenfarb, I. S., Cooper, L. D., & Grundt, A. M., (1999). The impact of acceptance versus control rationales on pain tolerance. *The Psychological Record, 49*, 33-48.

Hopkins, K. D., Hopkins, B. R. & Glass, G. V. (1996). *Basic statistics for the behavioral sciences* (3rd ed.). Needham Heights, MA: Allyn & Bacon.

Manly, B. F. J. (1997). *Randomization, Bootstrap, and Monte Carlo methods in biology* (2nd ed.). Boca Raton, Fl: Chapman & Hall.

Miller, W. S., & Armus, H. L. (1999). Directed forgetting: Short-term memory or conditioned response? *The Psychological Record, 49*, 211-220.

Ninness, H. A. C., McCuller G., & Ozenne, L. (2000). *School and behavioral psychology: Research in human-computer interactions, functional assessment, and treatment.* Norwell, MA: Kluwer Academic Publishers.

Ninness, H. A. C., Rumph, R., & Bradfield, A. (2001). Small group statistics for school psychologists: on-line randomization and permutation tests. *The Texas School Psychologist*, *8*, 11-12.

Ninness, H. A. C., Rumph, R., McCuller, G., Bradfield, A., Saxon, J., & Calliou, M. (2001). The effect of computer-emitted speech inflections during verbal-interactive responding. *The Psychological Record*, *51*, 561-570.

Peoples, M., Tierney, K. J., Bracken, M., & McKay, C. (1998). Prior learning and equivalence formation. *The Psychological Record*, *48*, 111-120.

Peres-Neto, P. R. & Olden, J. D., (2001). Assessing the robustness of randomization tests: Examples from behavioral studies. *Animal Behaviour, 61*, 79-86.

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

Romano, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *Annals of Statistics*, *17*, 141-159.

Roscoe, J. (1975). *Fundamental research statistics for the behavioral sciences.* New York: Holt, Rinehart, & Winston.

Sowell, E. J. (2001). *Educational research: An integrative approach.* Boston, MA: McGraw-Hill.

Savage, L. J. H. (1976). On rereading R. A. Fisher (with discussion). *Annals of Statistics*, *4*, 441-500.

Taylor, B. L. & Gerrodette, T. (1993). The use of statistical power in conservation biology: The vaquita and northern spotted owl. *Conservation Biology*, *7*, 489-500.

Todman, J. & Dugard, P. (2001). *Single-case and small-n experimental designs: A practical guide to randomization tests.* Mahwah, NJ: Lawrence Erlbaum Associates.

Williams, R. H., Zumbo, B. D., & Zimmerman, D. W. (2001). The scientific contributions of R. A. Fisher. *Edgeworth Series in Quantitative Educational and Social Sciences*, *7*, 1-23.

Winer, B.J. (1971). *Statistical Principles in Experimental Design* (2nd ed.). McGraw-Hill. New York.