Journal of Experimental Education, 62(1), 75–86

Relative Power of the Wilcoxon Test, the Friedman Test, and Repeated-Measures ANOVA on Ranks

DONALD W. ZIMMERMAN
Carleton University

BRUNO D. ZUMBO
University of Ottawa

ABSTRACT. Many introductory statistics textbooks in education, psychology, and the social sciences consider the Friedman test to be a nonparametric counterpart of repeated-measures ANOVA, just as the Kruskal–Wallis test is a counterpart of one-way ANOVA. However, it is known in theoretical statistics that the Friedman test is a generalization of the sign test and possesses the modest statistical power of the sign test for normal as well as many nonnormal distributions. Although not familiar to researchers, another significance test that can be regarded as a nonparametric counterpart of repeated-measures ANOVA is a rank-transformation procedure, in which the usual parametric statistical analysis is performed on ranks replacing the original scores. In the present computer simulation study we compared the ordinary paired-samples Student *t* test, the Wilcoxon signed-ranks test, and the sign test for correlated samples from normal, uniform, mixed-normal, exponential, Laplace, and Cauchy distributions, for which the relative efficiency of the methods is known. We also compared repeated-measures ANOVA, repeated-measures ANOVA on ranks, and the Friedman test for *k* mutually correlated samples from the same distributions, where *k* = 2, 3, and 4. Power functions revealed that the Friedman test performed like the sign test for all distributions, whereas ANOVA on ranks performed like the Wilcoxon test. These comparisons emphasize that classification of these statistical tests in introductory textbooks should be revised and that more attention should be paid to the rank transformation concept.

MOST INTRODUCTORY STATISTICS TEXTBOOKS written for educators, psychologists, and social scientists during the past several decades contain a chapter on nonparametric statistics. In recent years, the contents of that chapter have become remarkably uniform from one book to another. Statistical tests of location usually mentioned are (a) the Mann–Whitney–Wilcoxon test, which is a nonparametric counterpart of the Student *t* test; (b) the Wilcoxon matched-pairs signed-

ranks test, which is a counterpart of the paired-samples t test; (c) the Kruskal–Wallis test, which corresponds to the F test in a one-way analysis of variance (ANOVA) design; and (d) the Friedman test, which is considered to be a nonparametric alternative to repeated-measures ANOVA.

In the present article we replace this classification with one that is more harmonious with recent findings in mathematical statistics. We emphasize that the distinction between parametric and nonparametric methods hinges on the question of whether a sampling distribution of a test statistic is appropriate. Some results of a computer simulation study, which underscore this point of view and are incongruous with the standard textbook classification, are discussed.

The Friedman Test: A Generalization of the Sign Test

Authors of introductory statistics texts often convey the impression that the Friedman test is a generalization of the Wilcoxon test and is appropriate for many-sample designs in the same way that the Wilcoxon test is appropriate for two-sample designs. Furthermore, these authors lead one to believe that the Friedman test is a nonparametric alternative to repeated-measures ANOVA when assumptions underlying the F test are not satisfied, just as the Wilcoxon test is an alternative to the paired-samples t test. For example, Hays (1988) stated that "much as the Kruskal–Wallis test represents an extension of the Mann–Whitney test, the Friedman (1937) test is related to the Wilcoxon matched-pairs procedure" (p. 832). Similarly, May, Masson, and Hunter (1990) stated that "the rank test counterpart of the analysis of variance and randomization tests for k dependent samples is called the Friedman test" (p. 496).

It is known in theoretical statistics that the Friedman test is a generalization of the sign test and has the modest statistical power of the sign test for most distributions that are likely to be encountered in research (see, e.g., Conover, 1980; Noether, 1963; Puri & Sen, 1969). For normal distributions, the asymptotic relative efficiency (ARE) of the sign test with respect to the Wilcoxon test is .667. The ARE of the sign test with respect to the Student t test is .637.

Similarly, for normal distributions, the ARE of the Friedman test with respect to the F test is .955 $k/(k + 1)$, where k is the number of treatment groups (Noether, 1963; Sen, 1967). When $k = 3$, this value is .716. There is evidence that these asymptotic results are close to the relative efficiency to be expected for small and moderate sample sizes (Gilbert, 1972).

On the other hand, it has been discovered that the sign test is actually more powerful than both the Student t test and the Wilcoxon test for some nonnormal distributions, including the Cauchy distribution (see, e.g., Hodges & Lehmann, 1956; Randles & Wolfe, 1979). Given these facts, the Friedman test can hardly be regarded as a competitor of the F test for most distributions encountered in research. Certainly, the sign test is not ordinarily selected as an alternative to the t test when other choices are available. The Friedman test might be a competitor,

however, in the case of the Cauchy distribution and a few other nonnormal distributions.

A Rank Test Corresponding to Repeated-Measures ANOVA

From the above discussion, this question naturally arises: If parametric assumptions are not satisfied, what statistical test can replace repeated-measures ANOVA? Although the Friedman test is appropriate for some of the same experimental designs as repeated-measures ANOVA, the ARE values suggest that it is not as powerful an alternative as might be desired. But perusal of introductory texts reveals no well-known nonparametric test comparable in efficiency to the Wilcoxon test for these designs.

One solution to this problem is provided by the concept of a rank transformation (Conover, 1980; Conover & Iman, 1981). It is known that, for independent groups, the Mann–Whitney–Wilcoxon test is equivalent to an ordinary Student t test performed on ranks replacing the original scores (see also Nath & Duran, 1981). Similarly, the Wilcoxon signed-ranks test is equivalent to a paired-samples t test in which signed ranks replace difference scores. Finally, the Kruskal–Wallis test is equivalent to an F test performed on ranks replacing scores in a one-way ANOVA design. These findings imply that the distinction between parametric and nonparametric tests is more apparent than real.

This approach has led investigators to perform an ordinary repeated-measures ANOVA on ranks replacing the initial scores in the repeated-measures design. That is, sums of squares, mean squares, and so on, are calculated in the usual manner, the only difference being that ranks ranging from 1 to Nk replace the original scores. Similar methods have also been used in various factorial designs, discriminant analysis, and regression (see, e.g., Conover, 1980, for citations of many studies). Some other alternatives to repeated-measures ANOVA that are based on ranks have been proposed (e.g., Lehmann, 1975; Quade, 1979), and these are more comparable to the Wilcoxon test with regard to power. But authors of elementary texts still emphasize the Friedman test and rarely mention other possible choices. In the present article, we compared the rank-transformation procedure with the Friedman test and with the parametric F test performed on the initial scores.

Method

Computer Simulation of Paired-Samples and Repeated-Measures Designs

In a small computer simulation study the power of the significance tests mentioned above was compared for measures having normal, uniform, mixed-normal, exponential, Laplace, and Cauchy densities. These densities were chosen, not because of their relevance to practical research data, but because of their noticeable

influence on the power of the various significance tests. The already-known relative efficiencies of the t test, Wilcoxon test, and sign test in a paired-samples design provide a basis for comparison of the other procedures. All of the simulations were based on an experimental design with k treatment groups and N subjects in each group. In this study, k was 2, 3, or 4, and N was 14.

A computer program generated samples from pseudorandom numbers on the interval (0, 1); these samples were transformed to have distributions of a desired shape with known means and variances. Pairwise correlation of sample values in different treatment groups was induced by adding a multiple of one random variable to each of the random variables representing treatments. That is, if W, Y_1, Y_2, and Y_3 are independent with mean 0 and variance 1, then the correlation between $X_1 = (W + bY_1)(1 + b^2)^{-1/2}$ and $X_2 = (W + bY_2)(1 + b^2)^{-1/2}$ is ρ, where $b = [\rho/(1 - \rho)]^{1/2}$, and similarly for X_2 and X_3 and for X_1 and X_3. Throughout the present study the correlation was .60.

First, we performed three significance tests in a design with two treatment groups ($k = 2$). The tests were as follows:

1. *Paired-samples Student* t *test.* This test is essentially a one-sample Student t test performed on difference scores.
2. *Wilcoxon signed-ranks test.* This nonparametric rank test finds the differences between the initial scores, ranks these differences without regard to sign, and then affixes the sign of the difference to the rank. The test statistic is based on a sum of ranks and is equivalent to a paired-samples Student t in which a rank transformation is applied before the difference scores are entered into the t formula. This is the same as saying that the test is equivalent to a one-sample Student t test performed on signed ranks replacing the differences.
3. *Sign test.* This test uses the number of positive or negative differences as a test statistic. Because the information provided by ranking of differences is not considered, it is not surprising that the test is usually less powerful than the Wilcoxon test. For some distributions, however, discarding information can be beneficial. We have noted that for the Cauchy distribution the sign test is more powerful than the Wilcoxon test.

Next, we performed three significance tests in a repeated-measures design with two, three, or four treatment groups ($k = 2$, 3, or 4). The tests were as follows:

1. *Repeated-measures ANOVA.* This method is similar to one-way ANOVA using the F test, but the analysis allows for dependence among measures in the treatment groups by taking account of the interaction of a subject variable and the treatment variable in obtaining variance estimates. The paired-samples Student t test is a special case of the method, where $k = 2$, just as the independent-groups t test is a special case of one-way ANOVA, related by $t^2 = F$.
2. *Repeated-measures ANOVA on ranks.* In this method, the scores in all treatment groups are first combined into a single group and ranked. The resulting

ranks are then substituted for the original scores in their respective treatment groups, and the usual repeated-measures ANOVA is performed. There is no well-known nonparametric test that is equivalent to this method in the sense in which the Wilcoxon test is equivalent to a rank transformation applied to difference scores. We emphasize that this method is *not* equivalent to the Friedman test.

3. *Friedman test.* This test involves ranking scores within each of the treatment groups and calculating a test statistic from rank sums. As mentioned before, the Friedman test can be regarded as a generalization of the sign test. Its power slightly exceeds that of the sign test when there are more than two treatment groups but is not as great as the power of the nonparametric rank methods like the Wilcoxon signed-ranks test or the Mann–Whitney–Wilcoxon test.

In the present study, we applied all of the significance tests just described to samples from the following six population distributions.

1. *Normal distribution.* Normal deviates were generated using the method devised by Box and Muller (1958), which is based on the transformation $X = (-\log U_1)^{1/2}(\cos 2\pi U_2)$, where U_1 and U_2 are uniformly distributed pseudorandom numbers on the interval (0, 1). As a check on the method, the program also simulated normal deviates using two other methods: the Polar Marsaglia method, originated by Marsaglia and Bray (1964), and the central-limit method based on summing 12 uniformly distributed random numbers. Differences among the three methods were insignificant for purposes of the present study.

2. *Mixed-normal distribution* (or contaminated normal distribution). Samples were obtained from $N(0, 1)$ with probability .95 and from $N(0, 400)$ with probability .05.

3. *Exponential distribution.* This density was generated by $X = -\log U - 1$, where U is uniformly distributed.

4. *Laplace distribution* (or double-exponential distribution). This distribution was obtained by generating an exponentially distributed random variable X and transforming to $-X$ with probability .5.

5. *Cauchy distribution.* This heavy-tailed distribution, which does not have finite variance, was obtained from $X = \tan[\pi(U - .5)]$, where U is uniformly distributed.

6. *Uniform distribution.* This was a uniformly distributed pseudorandom number having mean 0 and variance 1.

Theoretical investigations, as well as computer simulations, have revealed that the paired-samples t test is more powerful than the Wilcoxon test for normal and uniform distributions, but that the reverse is true for mixed-normal, exponential, Laplace, and Cauchy distributions (Blair & Higgins, 1985; Hodges & Lehmann, 1956; Randles & Wolfe, 1979). For normal distributions, both the Wilcoxon test and the t test are more powerful than the sign test. For the Laplace distribution, the Wilcoxon test is the most powerful of the three, holding a slight advantage

over the sign test. However, in the case of the Cauchy distribution, the sign test is the most powerful, followed by the Wilcoxon test. More generally, it has been found in many investigations that parametric t and F tests are the most powerful for normal and uniform distributions, as well as various nonnormal distributions for which skewness and kurtosis are not extreme, while nonparametric methods often hold a power advantage for various heavy-tailed distributions.

Pseudorandom numbers were generated using a well-known and thoroughly tested prime-modulus multiplicative congruential generator described by Lewis and Orav (1989) and Lewis, Goodman, and Miller (1969). We found probabilities that test statistics exceeded critical values associated with the .05 significance level.

We introduced differences in location so that both Type I and Type II errors could be examined. In the case of the three-sample tests, we obtained differences in location by shifting scores in only one of the three treatment groups relative to the other two groups. There were 5,000 replications of the sampling procedure for each difference in location and for each distribution. For the two-sample simulations, significance tests were nondirectional.

Results

Table 1 contains the results of the simulations. Each entry in the table is the probability of rejecting H_0, based on 5,000 replications of the sampling procedure. In each section of the table, columns represent the various significance tests and rows represent effect size. For a difference of 0, the entry corresponds to the probability of a Type I error. Otherwise, the entry is 1 minus the probability of a Type II error, which can be regarded as the power of the test, provided the probability of a Type I error remains close to the .05 significance level.

From the first section of the table, representing the two-sample significance tests, it is apparent that the performance of the various tests is consistent with known ARE values and with results of previous simulation studies (compare, e.g., with Randles & Wolfe, 1979, p. 116). For all tests except the t test applied to the Cauchy distribution and the mixed-normal distribution, the probability of Type I errors remained close to the .05 significance level. When we introduced differences in location, the t test applied to the normal distribution and the uniform distribution slightly dominated the Wilcoxon test, which in turn dominated the sign test.

The pattern was reversed for the Cauchy distribution. In this case, the sign test and Wilcoxon test both were superior to the t test, and the sign test was superior to the Wilcoxon test. For the Laplace distribution and the exponential distribution, the Wilcoxon test was slightly more powerful than the t test, which in turn was slightly more powerful than the sign test. All the results were consistent with many previous studies of the relative efficiency of these tests.

TABLE 1
Probability of Rejecting H_0 by Paired-Samples *t* Test, Wilcoxon Text, and Sign Test

Distribution	δ	*t*	Wilcoxon	Sign
Normal	0	.051	.048	.055
	1	.196	.190	.171
	2	.552	.532	.446
	3	.878	.862	.753
Mixed-normal	0	.016	.042	.056
	1	.099	.200	.195
	2	.312	.558	.545
	3	.464	.799	.839
Exponential	0	.040	.048	.055
	1	.215	.246	.276
	2	.595	.637	.651
	3	.871	.887	.876
Cauchy	0	.018	.047	.056
	1	.067	.173	.228
	2	.194	.417	.557
	3	.335	.611	.767
Laplace	0	.047	.051	.056
	1	.194	.208	.211
	2	.570	.593	.557
	3	.873	.881	.834
Uniform	0	.053	.052	.059
	1	.179	.167	.149
	2	.543	.514	.403
	3	.877	.847	.700

Table 2 contains results of the various tests in the repeated-measures design. The pattern of results in Table 2, representing the two- and three-sample tests, as well as the many-sample tests applied to two treatment groups, was essentially the same as in the Table 1. In almost all instances, in fact, the probabilities of the corresponding two-sample and many-sample tests were close to each other. Further results not shown in the tables have been obtained using other sample sizes and numbers of treatment groups, and the pattern of results was essentially the same.[1]

Otherwise stated, the relative power of any pair of the many-sample significance tests in Table 2, expressed as a ratio of probabilities, is close to the relative power of the corresponding pair of two-sample tests in Table 1. The same is true for all six distributions. We conclude, therefore, that the Friedman test performs like the sign test for the normal distribution and the five nonnormal distributions

[1]Simulation results using other sample sizes and numbers of treatment groups can be obtained by writing to Donald W. Zimmerman, 2738 Garber St., Berkeley, CA 94705.

TABLE 2
Probability of Rejecting H_0 by *F* test, *F* test on Ranks, and Friedman Test in Repeated-Measures Design With *k* Groups

Distribution	δ	*F*	*F*(ranks)	Friedman
		k = 2		
Normal	0	.052	.051	.058
	1	.180	.170	.163
	2	.553	.504	.442
	3	.877	.838	.760
Mixed-normal	0	.020	.049	.059
	1	.103	.187	.203
	2	.315	.542	.543
	3	.465	.807	.841
Exponential	0	.038	.042	.055
	1	.204	.223	.278
	2	.589	.620	.647
	3	.875	.894	.882
Cauchy	0	.019	.048	.063
	1	.081	.173	.237
	2	.201	.445	.560
	3	.343	.695	.780
Laplace	0	.044	.047	.057
	1	.190	.191	.214
	2	.571	.573	.554
	3	.874	.872	.838
Uniform	0	.051	.049	.060
	1	.178	.169	.159
	2	.538	.478	.405
	3	.877	.812	.694
		k = 3		
Normal	0	.053	.051	.050
	1	.183	.166	.147
	2	.580	.537	.446
	3	.920	.887	.802
Mixed-normal	0	.015	.049	.048
	1	.077	.190	.189
	2	.242	.571	.578
	3	.378	.855	.881
Exponential	0	.044	.048	.050
	1	.201	.226	.259
	2	.617	.642	.657
	3	.899	.918	.893
Cauchy	0	.017	.042	.044
	1	.049	.153	.219
	2	.140	.437	.578
	3	.267	.732	.805

(Continued on next page)

TABLE 2 —(Continued)

Distribution	δ	F	F(ranks)	Friedman
Laplace	0	.048	.055	.051
	1	.192	.202	.198
	2	.605	.607	.578
	3	.911	.910	.875
Uniform	0	.051	.047	.042
	1	.168	.149	.125
	2	.570	.499	.399
	3	.916	.857	.736
		$k = 4$		
Normal	0	.049	.048	.045
	1	.174	.154	.126
	2	.584	.533	.447
	3	.929	.901	.823
Mixed-normal	0	.018	.049	.039
	1	.053	.170	.170
	2	.175	.565	.580
	3	.300	.865	.896
Exponential	0	.050	.052	.049
	1	.193	.227	.261
	2	.614	.646	.659
	3	.898	.917	.897
Cauchy	0	.018	.051	.047
	1	.038	.146	.207
	2	.108	.435	.581
	3	.215	.730	.816
Laplace	0	.049	.051	.043
	1	.174	.180	.179
	2	.602	.603	.583
	3	.917	.923	.900
Uniform	0	.050	.047	.043
	1	.162	.141	.123
	2	.580	.504	.397
	3	.931	.881	.757

that were examined, as far as probability of Type I and Type II errors is concerned. This result is consistent with the asymptotic results discussed earlier. Furthermore, the rank-transformation version of repeated-measures ANOVA performs like the Wilcoxon test for the same distributions.

Discussion

We examine these findings from a somewhat different perspective. It is possible to write one relatively simple computer program as a replacement for several

somewhat more complicated programs that are needed for various experimental designs described in introductory textbooks. This unified program will perform tests that are maximally efficient for data that might realistically be encountered in research. The program is just an algorithm for performing repeated-measures ANOVA for k treatment groups and N subjects in each group.

To analyze data from a simple paired-samples or matched-pairs design, for which a paired-samples t test is appropriate, one sets $k = 2$. And in order to handle data from heavy-tailed nonnormal distributions for which the Wilcoxon test is appropriate, one first transforms the scores to ranks and then sets $k = 2$. The same program, of course, will perform repeated measures ANOVA with $k = 3$, $k = 4$, and so on. Finally, it can analyze data for repeated-measures designs involving samples from heavy-tailed nonnormal distributions, by first transforming scores to ranks and then letting k be 3, 4, and so on.

The program will not yield maximal performance for samples from a Cauchy distribution, or a few other distributions, where the sign test is the best choice if $k = 2$ and the Friedman test is superior if $k \geq 3$. But perhaps this is not a serious shortcoming, because Cauchy distributions do not turn up frequently in educational and psychological research. The main decision of researchers in using this program is whether nonnormality is serious enough to justify transforming scores to ranks.

Let us now consider independent samples, where similar reasoning applies. A simple one-way ANOVA program can handle much of the elementary statistical analysis needed in research using such designs. That is, the independent-groups Student t test can be performed by setting $k = 2$, one-way ANOVA by setting $k = 3$, and so on. If normality is violated, one first transforms scores to ranks. Then, when $k = 2$, the program yields the same result as the Mann–Whitney–Wilcoxon test, and when $k \geq 3$ the same result as the Kruskal–Wallis test.

Focusing attention on the transformation to ranks, instead of the nonparametric test statistic, emphasizes the relationship of the tests to other procedures based on transformations. Researchers are familiar with various normalizing transformations, including square-root, logarithmic, and reciprocal transformations that have been used widely (see, e.g., Kirk, 1982, pp. 79–84, for a discussion of these techniques). All distributions of ranks are rectangular in shape, and, as far as power is concerned, conversion of data to rectangular form has advantages similar to conversion to normal form (Zimmerman & Zumbo, 1993).

Methods of robust estimation, which reduce the influence of outliers (see, e.g., Hampel, Ronchetti, Rousseeuw, & Stahel, 1986), provide still another approach to violation of assumptions. Outliers are implicated in many of the heavy-tailed distributions mentioned above. In some instances, methods based on detecting and downweighting outliers are more effective than rank methods. Robust estimation and the rank transformation have a similar advantage: A researcher need not ask, "Should a parametric or nonparametric test be used?" and then search for an appropriate test. Rather, the researcher asks, "Should this data be modified or

transformed in some way?" To answer the latter question, attention must be paid to the shape of the probability distribution of the measures that have been obtained. But once a decision is made about ranking or other modification, the significance testing procedure can proceed as usual without switching to different test statistics or different tables of critical values.

The rank-transformation approach has another advantage. For small sample sizes, it is sometimes not possible to perform nonparametric tests at the conventional .05 or .01 significance levels because of the discrete sampling distribution of test statistics based on rank sums. For example, when the Friedman test is performed, the scores within each treatment group are separately ranked. In the rank-transformation method, on the other hand, the scores from all treatment groups are first combined and then ranked together.

There are $(N!)^k$ ways of ranking k sets of N scores separately, but there are $(kN)!$ ways of ranking the combined kN scores all together, which is a considerably larger number. This means that there are more natural α-levels available for an F test on ranks than for a Friedman test; the availability of more natural α-levels can be a convenience if N and k are relatively small (see, e.g., Randles & Wolfe, 1979, p. 122, for a discussion of natural α-levels).

Because of the developments in robust estimation and inference mentioned above, as well as the equivalences of test statistics we have been considering, the distinction between parametric and nonparametric methods has become somewhat blurred. Current evidence suggests that the distinction, if it is to be useful at all for researchers in education, psychology, and social sciences, should be confined to the restricted mathematical framework in which it was originally formulated.

In the past, on the contrary, discussion of nonparametric tests in introductory statistics textbooks has been closely related to theories of scales of measurement. It is widely believed that parametric significance tests should be replaced by nonparametric methods when the underlying scale of measurement is an ordinal scale and not an interval scale. However, this point of view is difficult to reconcile with the equivalences discussed in the present article. If the initial data of a research study happens to be in the form of ranks, which is surely ordinal measurement, it makes no sense to insist on replacement of a parametric statistical test with an equivalent nonparametric test. See also Velleman and Wilkinson (1993) and Zumbo and Zimmerman (1993) for a recent discussion of misconceptions surrounding nominal, ordinal, interval, and ratio typology.

When faced with a serious violation of normality, a researcher sometimes can minimize Type I and Type II errors by replacing the initial measures by their ranks. Other types of robust inference, which sometimes are superior, are becoming available and more widely known. But handling violations by such methods is quite different from grappling with the more complex question of whether or not a certain level of psychological measurement has been achieved. The latter question, although having considerable theoretical interest for many investigators,

lies outside the context of probability and statistics in which the parametric–nonparametric distinction is meaningful.

NOTE

This research was made possible by a Carleton University Research Grant to the first author and by a Social Sciences and Humanities Research Council of Canada Fellowship at Carleton University to the second author.

REFERENCES

Blair, R. C., & Higgins, J. J. (1985). Comparison of the power of the paired samples *t* test to that of Wilcoxon's signed-ranks test under various population shapes. *Psychological Bulletin, 97,* 119–128.

Box, G. E. P., & Muller, M. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics, 29,* 610–611.

Conover, W. J. (1980). *Practical nonparametric statistics* (2nd ed.). New York: Wiley.

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician, 35,* 124–129.

Gilbert, R. O. (1972). A Monte Carlo study of analysis of variance and competing rank tests for Scheffe's mixed model. *Journal of the American Statistical Association, 67,* 71–75.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions.* New York: Wiley.

Hays, W. L. (1988). *Statistics* (4th ed.). New York: Holt, Rinehart & Winston.

Hodges, J., & Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the *t* test. *Annals of Mathematical Statistics, 27,* 324–335.

Kirk, R. E. (1982). *Experimental design* (2nd ed.). Monterey, CA: Brooks/Cole.

Lehmann, E. L. (1975). *Nonparametrics: Statistical methods based on ranks.* Holden-Day: San Francisco.

Lewis, P. A. W., Goodman, A. S., & Miller, J. M. (1969). A pseudorandom number generator for the System 360. *IBM Systems Journal, 8,* 136–146.

Lewis, P. A. W., & Orav, E. J. (1989). *Simulation methodolgy for statisticians, operations analysts, and engineers* (Vol 1). Pacific Grove, CA: Wadsworth.

Marsaglia, G., & Bray, T. A. (1964). A convenient method for generating normal variables. *SIAM Review, 6,* 260–264.

May, R. B., Masson, M. E. J., & Hunter, M. A. (1990). *Application of statistics in behavioral research.* New York: Harper & Row.

Nath, R., & Duran, B. S. (1981). The rank transform in the two-sample location problem. *Communications in Statistics: Simulation and Computation, 10,* 383–394.

Noether, G. E. (1963). Efficiency of the Wilcoxon two-sample statistic for randomized blocks. *Journal of the American Statistical Association, 58,* 894–898.

Puri, M. L., & Sen, P. K. (1969). On the asymptotic theory of rank order tests for experiments involving paired comparisons. *Annals of the Institute of Statistical Mathematics, 21,* 163–173.

Quade, D. (1979). Using weighted rankings in the analysis of complete blocks with additive block effects. *Journal of the American Statistical Association, 74,* 680–683.

Randles, R. H., & Wolfe, D. A. (1979). *Introduction to the theory of nonparametric statistics.* New York: Wiley.

Sen, P. K. (1967). A note on the asymptotic efficiency of Friedman's test. *Biometrika, 54,* 677–79.

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *American Statistician, 47,* 1–7.

Zumbo, B. D., & Zimmerman, D. W. (1993). Is the selection of statistical methods governed by level of measurement? *Canadian Psychology, 34,* 390–400.

Zimmerman, D. W., & Zumbo, B. D. (1993). The relative power of parametric and nonparametric statistical methods. In G. Keren, & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 481–517). Hillsdale, NJ: Erlbaum.