

SPECIAL SECTION: STATISTICAL ANALYSES FOR SMALL SAMPLES

**MULTIVARIATE RANDOMIZATION TESTS FOR SMALL-N
BEHAVIORAL RESEARCH: A WEB-BASED APPLICATION**

Chris Ninness, Robin Rumph, Eleazar Vasquez III, and Anna Bradfield¹
Stephen F. Austin State University

Sharon K. Ninness
Angelina College

ABSTRACT: Behavioral research involving statistical analyses of small group data is frequently compromised by conventional parametric statistical procedures. As an alternative, we have developed and deployed several web-based applications that allow behavioral researchers to easily input data on-line and to calculate levels of significance for small-*n* studies. Previously, our web-based applications were restricted to parametric and randomization tests involving only one dependent variable. We now have expanded our algorithms such that multivariate analyses may be conducted on sample sizes as small as 6 while employing several dependent measures. This paper provides details for on-line input and interpretation of randomized multivariate statistical tests for small-*n* studies. Also, to test the power and reliability of our applications, we have compared our multivariate randomization algorithms against a traditional multivariate statistic with two dependent variables. Using Monte Carlo methods, we have assessed the statistical advantages and accuracy of applied multivariate analyses when calculated in both randomized and traditional/parametric formats. Specifically, we have compared probability values for both traditional and randomized MANOVA models by way of *Hotelling's T²* and the *Randomized multivariate/composite z scores*. We discuss the reliability problems associated with using traditional multivariate statistics with small-*n* studies, and we describe the statistical advantages and some limitations of using our on-line, small-*n*, multivariate randomization tests.

Sidman (1960) has pointed out that in traditional psychological research, subject variability is considered a source of experimental *error* while in behavioral research, it is a source of experimental *interest*. There can be no question that in well-controlled single-subject preparations, the record of response variability is sufficient to demonstrate that behavior change is (or is not) a function of treatment. Although behavioral research always will be grounded in single-subject design, it is not unusual for applied (e.g., Kollins, Lane, & Shapiro, 1997) and basic (e.g., Theodoratus, Chiszar, & Smith, 1997) researchers to employ small group experimental investigations. Recently, it has become fairly common for behavioral researchers to examine more than one dependent measure and to analyze data

¹ Funding for www.lcsdg.com/psychstats was provided by a Texas Excellence Funds Grant through Stephen F. Austin State University. We gratefully acknowledge the special help and guidance provided by Morris Lang in the development of this web-based project. We are also extremely thankful to Jeremy Cobb and Michael Coffee for their tireless efforts in the refinement of this work. Correspondence concerning this article should be directed to Chris Ninness, School & Behavioral Psychology Program, Department of Human Services, PO Box 13019 SFA Station, Stephen F. Austin State University, Nacogdoches, TX 75962 (e-mail: cninness@titan.sfasu.edu).

using traditional parametric univariate tests. However, with increasing sample sizes and multiple dependent variables, control over extraneous variables becomes very unreliable. Particularly when small- n measures are taken and several dependent variables are analyzed using traditional univariate statistics, the internal validity of the study is seriously compromised.

Consider a study in which five dependent measures were assessed concurrently, and separate traditional univariate t -tests were performed on each dependent variable. Here, multiple dependent variables were analyzed separately in univariate fashion—each separate analysis failed to take the other into account. Unfortunately, this is by no means an unusual occurrence (See Stevens, 2001, for a complete discussion). Under these conditions, the use of fragmented univariate t -tests inflate the overall type I error rate, and the investigators conclude that treatments were effective in conditions where pure chance might well have been operating. The probability of one or more spurious results when five t -tests are employed simultaneously is substantially above the .05 type I error rate (See Stevens, 2001, for a discussion). With 5 concurrent t -tests, the probability of *no* type I errors is: $(.95)(.95)(.95)(.95)(.95) = .774$, since the chances of *not* making a type I error for each test is .95. In this example, the likelihood of making at least one type I error is $1 - .774 = .226$. It is possible that appropriate multivariate and *post hoc* tests (or Bonferroni adjustment) might have revealed statistical significance among several of the dependent measures; however, researchers should not make such a claim until they have analyzed their data correctly. Adding to these complications, the above hypothetical study employs a relatively small sample size. Roscoe (1975) and Siegel and Castellan (1988) suggest that in the absence of an unambiguous demarcation between large- n and small- n studies, there is a commonly held assumption among parametric statisticians. To paraphrase the perspective of most authorities, *the smaller the group size, the harder it is to be certain that the normal curve assumptions have not been violated* (Todman & Dugard, 2001).

Even if an appropriate MANOVA test had been employed in the above scenario, Monte Carlo methods have shown that traditional MANOVA tests have insufficient power to identify the significant differences that actually exist with small- n data (Chen, 1993). As with univariate statistical tests, it has been widely reported that the power and reliability of the parametric multivariate procedures dwindle as the sample size shrinks (e.g., Davidson, 1972). This sort of grim statistical prognosis for small- n research detailed in most multivariate research and textbooks has dissuaded many behavior analysts from even considering any form of multivariate computation as a viable option. However, when behavior analysts look to statisticians for alternatives to traditional univariate or multivariate strategies, various nonparametric options such as Mann-Whitney U and Wilcoxon T are recommended; unfortunately, these procedures have notoriously weak sensitivity to treatment effects with small- n data (Todman & Dugard, 2001).

But behavior analysts are far from being alone when it comes to statistical enigmas. There is an amazing irony in the continuing failure of modern researchers to confront the inescapable fact that the vast majority of experimental research

generated at universities is conducted with nonrandom and inadequate sampling procedures. In the apparent absence of any alternative “logic,” the tradition of gathering large but “restricted samples” and making inferences to much wider populations remains a subtle but inane standard throughout academia. Todman and Dugard (2001) draw our attention to the inexorable fact; “It is difficult to conclude other than that random sampling in human experimental research is little more than a convenient fiction. In reality, generalization almost invariably depends on replication and nonstatistical reasoning”(p. 4).

The hard reality is that very few so called “large group studies” in psychology, and particularly studies involving human behavior, even approximate fulfilling the assumptions of true random sampling procedures that require each and every element in the entire “relevant population” to have an equal and independent opportunity of being selected for participation (Edgington, 1995). More often, when subjects are identified for traditional statistical analyses, they are drawn randomly from the population at a given university. The external validity is further weakened by employing subjects who are *willing to participate to partially fulfill course requirements*. It is common knowledge that few experimental, clinical, or social psychologists (and a very wide range of related academics) have no reservations about using traditional hypothesis testing procedures based on normal curve distributions when running fifty subjects using such unrepresentative formats (Edgington, 1995; Good, 1994). Somehow, it is widely assumed that a larger number of conveniently located and immediately available subjects from a given location may compensate for the lack of a broader and more representative but inconveniently located group of representative participants (e.g., Hastings, Remington, & Hall, 1995). Such sampling procedures are based on the assumption that fifty rats or sophomores at one university may be very much like fifty rats or sophomores at another. Depending on the variables being tested and the group being treated, such logic may be perfectly correct or completely erroneous, but it has nothing to do with statistical inference based on normal curve theory. Rather, the external validity can only be based on *logical probability*.

However, there is an alternative approach that, up until recently, has eluded the attention of much of the academic community. There exists a body of reliable and precise algorithms that do not entail any underlying assumptions regarding random and independent sampling, homoscedasticity, or normal distribution of the parent population. Randomization tests (Edgington, 1995) are a series of precise computational procedures generating probabilities based on all possible permutations that could occur with a given data set (see Good, 1994, for a similar discussion on permutation tests). Using these conservative but extremely accurate algorithms, a researcher who computes statistical significance of a small-*n* study need not be concerned that the assumptions underlying the normal curve have somehow been violated. The randomization test computes all possible permutations for any number of data points and calculates the probability that the obtained results could have occurred simply as a matter of chance. It is important to recognize the one essential feature for maintaining the *internal validity* of randomization tests. If between group designs are employed, these computations

can only be run accurately if there is random assignment of subjects (or observations) to treatments (Efron & Tibshirani, 1993). (Alternatively, univariate *randomized repeated measures* and *randomized correlated t-test* are available and do not entail random assignment of subjects to groups, since the same subjects are employed in sequential measures of the dependent variable.) Unlike parametric statistics, the *external validity* is not predicated on normal curve assumptions underlying the required numbers at various degrees of freedom within the significance tables. Precise probability values (P-values) are calculated by randomization tests, but the external validity of obtained P-values can only be judged by evaluating the logical probability that other populations share germane characteristics of the participants within a given study.

Historically, randomization tests have not been popularly employed because their extensive computations required more processing speed than the existing technology could provide, but this has changed dramatically in the last few years. Recently, Peres-Neto and Olden (2001) have shown that randomization tests calculate very sensitive type I error rates for univariate procedures when contrasted against the traditional parametric ANOVA and the nonparametric Kruskal-Wallis test. Indeed, this research shows that randomization tests appear to be generally more robust to violation of the statistical assumptions associated with parametric and classic nonparametric approaches. However, Peres-Neto and Olden emphasize that one should not assume, a priori, that randomization tests are “always” more robust (i.e., type I error and power) than other methods. They suggest that explicit comparison of methods using the data at hand is the most advantageous approach. Even so, very little research has been conducted on the development of *multivariate randomization tests*. For example, Chung and Fraser (1958) provided an algorithm for reducing the total number of permutations by having the computer systematically select a subset of the data on which the probabilities of all permutations could be estimated. Using this general strategy to obtain approximated probabilities, these researchers proposed a test statistic for MANOVA; however, the technique has not been widely accepted or employed. As previously noted, computer processing speed has been insufficient to allow access by potential users, and even among those who espouse the potential advantages of the various randomization tests, *multivariate* randomization procedures have not been developed and tested for applied research (see Manly, 1991, p. 264, for a detailed discussion).

In fairness, Willmes (1988) produced an early Fortran multivariate test using Pillai’s trace test statistic; however, this procedure is not directly accessible on the web. Chen (1993) developed an “approximate” randomization test for several MANOVA procedures in Fortran. Results were encouraging but inconclusive because the software and machine processing speed only generated a “sub-sample” of all possible permutations to “estimate” levels of significance. Outcomes based on this approximate randomization test by Chen suggest that for small sample sizes, there appears to be a substantial gain in statistical power when employing *Randomized multivariate/composite z scores*.

More recently, *univariate* randomization tests have started to gain some recognition in applied and basic research across a wide range of academic disciplines (e.g., Johnson & Johnson, 1989; Mundry, 1999, Thomas & Poulin 1997). Nevertheless, due to the lack of information regarding the performance of *multivariate* randomization tests, that particular statistical option has not been made available to applied or basic researchers.

Extending the algorithms originally described by Edgington (1995), we have constructed and deployed an easy-to-use series of univariate and multivariate randomization tests. These small-*n* statistical procedures are freely available on-line to all academic researchers. In order to examine a wide range of completely diversified outcomes computed by our multivariate randomization test, we used a Monte Carlo method to generate small-*n* data points on two dependent variables. Significance tests for group differences were computed by way of the traditional parametric statistic *Hotelling's T*² and by the *randomized composite z-scores* algorithms. The respective probability values produced by these procedures were depicted graphically and compared.

METHOD

Apparatus and Software

The research was conducted on a Dell Dimension 4400 computer (Pentium [4] 1.8 GHz processor with 768 MB RAM). Statistical software was written by the first author in Visual Basic 6 and C++ for IBM PC compatible machines; however, the original *randomization test* algorithms were developed in Fortran by Edgington (1995). These algorithms were adapted and extended to run in the C++ language as developed and described by Ninness, Rumph, & Bradfield (2001). All algorithms were audited and refined by LCSDG, LLC (a university-affiliated computer consulting firm). 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

To evaluate the effects of small sample size on MANOVA levels, we employed Microsoft's pseudorandom number generator within our C++ code to produce distributions of 1000 virtually random values ranging from 00 to 99 for samples with *n*'s of 12, 10, 8, and 6 (on two dependent variables). Thus, our data points were not derived from any preexisting simulated distributions (e.g., skewed or bimodal). Using the computer's clock, each file (composed of *n* random values) was individually labeled in nanoseconds at the moment it was generated. Following the production of these data sets, randomized multivariate/composite *z-scores* and the traditional parametric *Hotelling's T*² were computed on each Monte Carlo data set. P-values for each data set were compared and depicted graphically.

Bonferonni adjustment. Conventional multivariate procedures, including *Hotelling's T*², yield probabilities that are composed of linear combinations of separate dependent measures. Therefore, when calculating *Hotelling's T*², one of

the two dependent measures may show a significant difference between groups (the discriminative function), while the second measure fails to do so. This is an especially valuable feature of Hotelling's T^2 ; however, as pointed out by Stevens (2001), a very large sample size is needed to reliably identify any "meaningful variate" when employing this procedure. Conversely, the randomized composite z -score does not attempt to provide any form of discriminative function whatsoever. Rather, the composite z -score represents a unified estimate of the combined effects on both dependent measures—simultaneously. Any comparison of statistical probabilities generated by the respective models must take this into account. Thus, a Bonferroni adjustment was employed on all obtained composite z -scores. For composite z -scores, our software multiplied each obtained P-value by the number of dependent variables (in this case 2) to represent the corrected experimentwise error rate.

RESULTS

Figure 1 shows the scatterplot, correlation, regression, and percentages of P-values that fell at or below .05 for group sizes of six in each group as calculated by way of Hotelling's T^2 and randomized composite z -scores procedures. In the first Monte Carlo generation of 1000 data sets of size 6, P-values for randomized composite z -score tests and parametric Hotelling's T^2 correlated at .93. Here, 6.3% of the obtained parametric probabilities fell at or below the truncated .05, while only 4.4% of the randomization test P-values were at or below this level.

Figure 1 also shows the same information for n 's of 12 with a conspicuous improvement in the amount of scatter along the regression line. P-values correlated at .95, and 5.6% of the Hotelling's T^2 outcomes were at or below the designated level at .05. The randomized composite z -scores showed 4.8% of the P-values below .05. In both conditions, the randomized composite z -scores P-values that fell at or below .05 always corresponded very closely to those identified by the parametric Hotelling's T^2 ; however, these parametric P-values were consistently more *liberal* than the composite z -score and often exceeded the specified .05 level (see circled areas on Figure 1).

Because the scatterplots produced from group sizes 8 and 10 showed essentially the same trends, they are not displayed; however, the correlations between parametric and randomized t -test probabilities were found to be .94 and .95, respectively. Moreover, in the randomized composite z -scores, P-values were slightly (but consistently) more conservative than those obtained by Hotelling's T^2 . We replicated all of the above Monte Carlo data sets for sample sizes 12, 10, 8, and 6, and the obtained correlations were all within .0001 of our first series. Likewise, the percentage of P-values falling at or below .05 in each case was virtually identical (within .001).

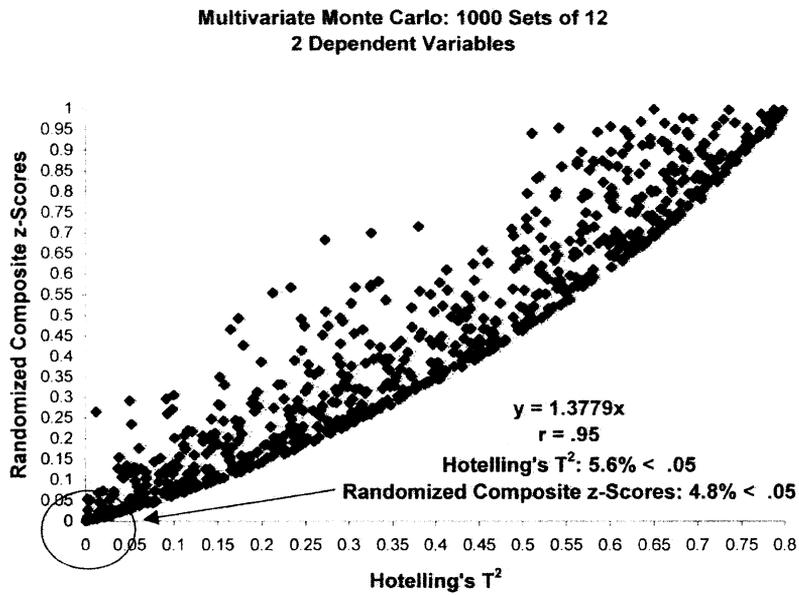
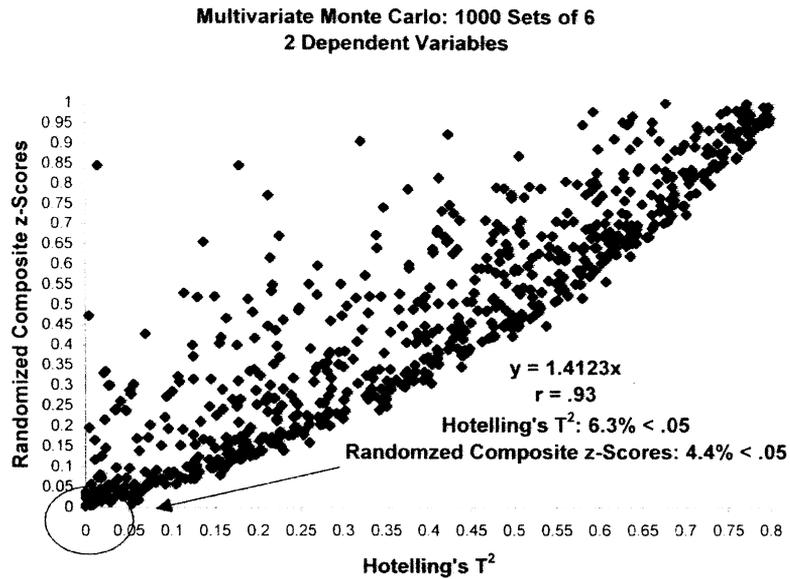


Figure 1. Scatterplot, correlation, regression, and percentages of P-values that fell at or below .05 for group sizes of 6 and 12 as calculated by way of Hotelling's T^2 and randomized composite z-scores procedures.

DISCUSSION

Our recent research (Ninness, Newton, Saxon, Rumph, Bradfield, Harrison, Vasquez, & Ninness, 2002) has shown the relatively strong correspondence between classic and randomized statistical strategies—at least in terms of the commonly employed two-sample independent t -test with equal n 's. In this study and the previous univariate study (Ninness et al., 2002), P-values that allowed rejection of the null hypothesis (set at .05) fell in some disagreement as group sizes shrank from 12 to 6 in steps of 2. Even with these relatively small n 's, the correlations between Hotelling's T^2 and randomized composite z -scores were at .95, .95, .94, and .93, for group sizes of 12, 10, 8, and 6, respectively. Interestingly, the randomized composite z -score P-values almost always corresponded with those generated by the parametric Hotelling's T^2 , but the inverse was not true when sample sizes were less than 12. With smaller n 's, the parametric P-values gradually became inflated above the designated level while the randomized composite z -score tests remained stable, finding slightly less than 5% of the P-values below the set at .05. Thus, with n 's less than 12, it appears that the randomized composite z -scores test consistently represents a more conservative but reliable measure of type I error rates. For sample sizes of 12 or more in each group, there appears to be very little difference in P-values obtained by either approach. To some extent, our results may be attributable to the relatively uniform distributions created by our Monte Carlo method. Had we used exponentially distributed data or log-normal data, the relationships with Hotelling T^2 might not have been so strongly correlated (B. Manly, personal communication, November 29, 2002). Our future research will explore this and related Monte Carlo method possibilities.

Statistical inference based on randomization test P-values is another matter. As noted at the beginning of this paper, it is very rare for traditional large group research involving human behavior to satisfy the primary assumptions underlying the specified levels for making inferences to populations based on samples (Todman & Dugard, 2001). This is not to suggest that specific P-values obtained by way of parametric statistics are incorrect. Rather, it is the parametric inference from non-representative samples to wider populations that is gratuitous. As noted by Edgington (1995), Manly (1991), Hopkins et al. (1996), and many others, the size of a given sample is the least essential feature of any statistical analysis. Unless it is very clear that a sample is *randomly* drawn from the very population it is supposed to represent, the external validity of a *large group* study cannot be inferred from the sample. External validity can only be addressed by judging the *logical probability* that other populations share the germane characteristics of the individuals who did not participate in a given study. This is the obvious position assumed by single-subject operant researchers; it is the natural position assumed by those who employ randomization tests. According to Hopkins et al. (1996), it should be the position adopted by any researcher who employs human participants and attempts to calculate significance based on traditional parametric statistical tables.

On-line Interactive Data Analysis

Our website at www.lcsdg.com/psychstats has been developed within the Stephen F. Austin State University School & Behavioral Psychology Program and is free to all academic users. Initially, the user must set up his/her user name and password with the new account button. Subsequently, the on-line software provides immediate access and facilitates straightforward selection of the various types of statistical tests (e.g., Univariate, Multivariate, Univariate Nonparametric, Multivariate Nonparametric) for a wide range of data types (including binary). The Multivariate Randomization Tests become available after the user clicks the *Multivariate Nonparametric in C++* button. The functions on this page simply ask the user to input the number of subjects in each group and the number of dependent variables. The on-line software automatically generates the needed spreadsheets and guides the user through input of all raw scores. After entering the data, the user clicks "Finish," and the program immediately computes the statistical probabilities. (Note: The conventional univariate and multivariate pages provide probability values in conjunction with related statistical calculations.) The results can be printed, saved, or copied directly from the site.

As Edgington (1995) points out, the process of computing composite z -scores by combining z -scores is predicated on the assumption that the multiple dependent measures change in the same direction. That is, both dependent measures have the same valence. If treatments are designed such that the effects may be in opposing directions, the signs of associated z -scores must be reversed. For example, if we develop an intervention for six students aimed at *increasing* their scores on a math test and *decreasing* the time it takes them to complete the test, the respective outcomes should show positive effects that go in opposite directions. Following treatment, students show *higher* scores and take *less* time to complete the test. Quantitatively, this matter is easily resolved by reversing the signs of data points relative to the time variable. Important to note is that following a finding of family-wise significance with the *randomized multivariate analysis*, post hoc randomized t -tests can be run on the *psychStats* server.

There are several caveats regarding the use of *psychStats*. Our application server is not designed to function as a "full-service" statistical package. While the on-line functions provide immediate results on a large number of conventional and randomization tests, data cannot be stored on our server. Data entry entails inputting all values each time a given statistical test is employed. The server will not store data files after the user has completed an analysis, and it will not allow the user to load large data files from existing spreadsheets. The functions on *psychStats* were developed to accommodate small- n experimental design procedures. Also in this regard, it is important to point out that covariance is not calculated in multivariate randomization tests. However, even if this were not the case, covariance is not a very useful measure with any form of small group data analysis (Stevens, 2001).

Presently, the multivariate randomization tests on the *psychStats* server can reliably compute the probability values for several parametric, nonparametric, and

randomization tests. For parametric and randomized multivariate tests, the applications will reliably compute group differences based on two and three dependent variables (Vasquez, 2002). New algorithms for *randomized multivariate repeated measures* (see Edgington, 1995, for a discussion), as well as variations on *artificial neural networks* approaches (e.g., Kohonen, 2001) adapted for repeated measures designs, are under development in our laboratory. Hopefully, these will serve as useful adjuncts to present on-line multivariate architecture.

REFERENCES

- Chen, R. (1993). *Approximate randomization tests for Manova*. Dissertation Abstracts: Tulane University.
- Chung, J. H., & Fraser, D. A. S. (1958). Randomization test of a multivariate two-sample problem. *Journal of the American Statistical Association*, 53, 729-735.
- Davidson, M. L. (1972). Univariate versus multivariate tests in repeated-measures experiments. *Psychological Bulletin*, 77, 446-452.
- Edgington, E. S. (1995). *Randomization tests*. New York: Marcel Dekker, Inc.
- Efron, B. & Tibshirani, R. J. (1993). *An introduction to the Bootstrap*. New York: Chapman & Hall.
- Glass, G. V., Peckham, P. D., & Sanders., J. R. (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 resampling methods for testing hypotheses*. New York: Springer-Verlag.
- Hastings, R. P., Remington, B., & Hall, M. (1995). Adults' responses to self-injurious behavior: An Experimental Analysis Using a Computer-Simulation Paradigm. *Behavior Modification*, 19, 425-450.
- Hopkins, K. D., Hopkins, B. R., & Glass, G.V. (1996). *Basic statistics for the behavioral sciences* (3rd ed.). Needham Heights, MA: Allyn & Bacon.
- Hotelling, H. (1931). The generalization of student's ratio. *Annals of Mathematical Statistics*, 2, 360-378.
- Johnson, D. W., & Johnson, R. T. (1989). *Cooperation and competition: Theory and research*. Edina, MN: Interaction Book Company.
- Kohonen, T. (2001). *Self-Organizing maps: 2nd Edition*. Springer Series Information Sciences,
- Kollins, S. H., Lane, S. D., Shapiro, S. K. (1997). Experimental analysis of childhood psychopathology: a laboratory matching analysis of the behavior of children diagnosed with attention-deficit hyperactivity disorder (ADHD). *The Psychological Record*, 47, 25-44.
- Manly, B. J. F. (1991). *Randomization and Monte Carlo Methods in Biology*. Chapman and Hall:London.
- Mundry, R. (1999). Testing related samples with missing values: A permutation approach. *Animal Behaviour*, 58, 1143-1153.
- Ninness, H. A. C., McCuller G , & Ozenne, L. (2000). *School & behavioral psychology: Research in human-computer interactions, functional assessment, and treatment*. Norwell, MA: Kluwer Academic Publishers.
- Ninness, H. A. C., Newton, R., Saxon, J., Rumph, R., Bradfield, A., Harrison, C., Vasquez, E. & Ninness, S. (2002). Small group statistics: A Monte Carlo comparison of parametric and randomization tests. *Behavior and Social Issues*, 12, 53-63.

- 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*, Spring, 11-12.
- Peres-Neto, P. R., & Olden J. D. (2001). Assessing the robustness of randomization tests: examples from behavioural studies. *Animal Behaviour* 61, 79-86.
- Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any populations. *Journal of the Royal Statistical Society*, 4, 119-130.
- Roscoe, J. T. (1975). *Fundamental Research Statistics for the Behavioral Sciences*. New York: Holt, Rinehart and Winston, Inc.
- Sidman, M. (1960). *Tactics of Scientific Research*. New York: Basic Books.
- Siegel S. & Castellan, N. J. (1988) *Nonparametric Statistics for the Behavioural Sciences* (2nd ed.). McGraw-Hill, New York.
- Stevens, J. (2001). *Applied Multivariate Statistics for the Social Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Theodoratus, D. H., Chiszar, D., Smith H.M. (1997). Rattlesnake Orientation to Prey as a Function of Thermal Backgrounds and Edges. *The Psychological Record*, 47, 461-472.
- Todman, J. & Dugard, P. (2001). *Single-case and small-n experimental designs: A practical guide to randomization tests*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Thomas, F. & Poulin, R. (1997). Using randomization techniques to analyze fluctuating asymmetry data. *Animal Behaviour*, 54, 1027-1029.
- Vasquez, E. (2002). *A comparison of parametric and randomized multivariate statistics using Monte Carlo methods: A computer generated approach with three dependent variables*. Poster session presented at the annual meeting of the Texas Association of School Psychologists Conference, Dallas, TX.
- Willmes, K. (1988) *Beitrage zu Theorie and Audwendung von Permutationstest in der Uni- and Mutivariaten Datenanalyse*. Ph.D. Thesis, University of Trier.