Journal of Educational Statistics
Winter 1983, *Volume* 8, *Number* 4, *pp. 289–309*

N = 1 DESIGNS: THE FAILURE OF ANOVA-BASED TESTS

LARRY E. TOOTHAKER, MARTHA BANZ, CINDY NOBLE, JILL CAMP, AND DIANA DAVIS
The University of Oklahoma

Key words: *N* = 1 *Designs*, *ANOVA-based tests*, *serial correlation*, *robustness*.

ABSTRACT. Several methods have been proposed for the analysis of data from single-subject research settings. This research focuses on the modifications of ANOVA-based tests proposed by Shine and Bower, a procedure that precedes the ANOVA *F* test by preliminary testing of within-phase lag one serial correlation and the one-way ANOVA as presented by Gentile, Roden and Klein. Monte Carlo simulation is used to investigate these tests with respect to robustness and power. Each test was analyzed under various patterns of serial correlation, various patterns of phase and trial means, normal and exponential distributions, and equal and unequal phase variances. The findings indicate that the probability of a Type I error for these ANOVA-based tests is seriously inflated by nonzero serial correlation. These tests, therefore, cannot be recommended for use with data that have nonzero serial correlation.

Single-subject research settings exist for a wide range of research topics in a variety of disciplines: community research (Friesema, Caporaso, Lineberry, & Goldstein, 1978), political science (Deutsch & Alt, 1977), law (Glass, Tiao, & Maguire, 1971), medicine (Sterman, 1973), behavioral research (Juliano & Gentile, 1974; Leitenberg, Agras, & Thomson, 1968), school and/or educational psychology (Gottman & McFall, 1972; Kratochwill, 1977), clinical psychology and psychiatry (Barlow & Hersen, 1973; Chassan, 1967), and experimental psychology (Shine, Wiant, & DaPalito, 1972). Treatment of data from such designs has been controversial in that some advocate visual inspection while others call for inferential methods (Kratochwill, 1978). In addition, controversy exists over choice of inferential methods, including advocacy of traditional statistical tests which assume independence of observations. Dangers in using traditional methods include excessive rejection rates for serially correlated data (Box, 1954; Hibbs, 1974) while dangers in visual inspection include inability to separate chance from nonchance variability and lack of clarity of decision rules (for discussion on visual inspection see Kazdin, 1982, and Parsonson & Baer, 1978). While single-subject designs need special consideration with respect to the question of validity (see Levin, Marascuilo, & Hubert, 1978), if researchers are interested in testing hypotheses in such designs, they must choose from several available methods.

Inferential methods that have been proposed for testing treatment or intervention effects in single-subject research include graphical analysis (Parsonson & Baer, 1978), modifications of traditional tests, which are usually based on the ANOVA (Shine & Bower, 1971), randomization tests (Edgington, 1972), Markov chain analysis (Gottman & Notarius, 1978), and time series (Box & Jenkins, 1976; Glass, Willson, & Gottman, 1975). This article will focus on the ANOVA-based modifications of traditional tests and the problems of such tests associated with serially correlated data.

Traditional statistical tests, such as *t* tests or ANOVA models on the observed data have been both recommended by applied researchers (Gentile, Roden, & Klein, 1972; Hedberg, Walker, & Ehrman, 1975) and condemned by methodological experts (Cook & Campbell, 1979). The latter opinion is based on using the traditional tests on data that are serially correlated and thus in violation of the assumption of independence of the residual or error component. The validity of these ANOVA procedures is based on independent errors. Whether or not single-subject data violate the independence assumption is also controversial, with some researchers claiming large positive lag-one serial correlations (Hartmann, Gottman, Jones, Gardner, Kazdin, & Vaught, 1980; Jones, Vaught, & Weinrott, 1977) and others claiming zero population lag-one serial correlation for behavior modification data (Huitema, 1983). Differences of view may depend on whether serial correlations are computed within phases or over the total data, and on area of the research. The very presence of the serial correlation and partial serial correlation patterns analyzed by time series analysis for single-subject data indicates a potential problem with meeting the assumption of independent errors. Traditional significance tests are biased positively (too many rejections) if the serial correlation between observations is positive. Box (1954) showed that if lag one serial correlation was $\rho = .4$, then $\alpha = .25$ for an ANOVA *F* test when the probability of Type I error was set at $\alpha = .05$. Gastwirth and Rubin (1971) showed for $\rho = .4$, the one-sample *t* test had $\alpha = .14$, the one-sample Wilcoxon test had $\alpha = .14$, and the one-sample sign test had $\alpha = .11$ (all when set $\alpha = .05$). However, these tests were not designed for single-subject data nor presented by these authors as appropriate for such data. Gentile et al. (1972) presented a test which is the usual one-way ANOVA *F* test on between-phase and within-phase variability, and suggested it could be used for $N = 1$ data in the ABAB reversal design and that serial correlation would work toward conservation in the *F* test. Examples of ANOVA-based modifications of traditional statistical tests designed or at least presented as appropriate for hypotheses in single-subject data include the ANOVA model of Shine and Bower (1971) and preliminary testing procedures such as testing for nonzero, within-phase, lag-one serial correlation and proceeding to the ANOVA *F* or *t* test only if all correlations are not significantly different from zero (see Kazdin, 1982). (For a more complex preliminary testing procedure, see Hartmann, 1974.) Shine and Bower (1971) presented what they called a "one-way" ANOVA, which was actually a modification of the simple repeated measures design (SRMD) or a two-way ANOVA with one observation per cell. Instead of using a group of I random subjects measured once for each of J fixed levels of an experimental factor (as is the case for the SRMD), the Shine-Bower model assumes a series of I fixed trials (A) for each of J fixed levels of an experimental factor (B) for only one subject. Tests for B and AB are given by

$$F_B = MS_B/MSE', \qquad df = J - 1, I/2,$$
$$F_{AB} = MS_{AB}/MSE', \qquad df = (J - 1)(I - 1), I/2,$$

where

$$MSE' = \frac{J}{2} \sum_{i=\text{odd}}^{I-1} (\bar{Y}_{i+1} - \bar{Y}_i)^2/I/2.$$

Successive differences of trial means are squared and summed over the odd values of i, from 1 to $I - 1$ (for even I). MSE' is supposed to have expected value of σ_e^2 if the effect of trial i is equal in the population to the effect of trial $i + 1$ for odd i (this is the slow change assumption given by Shine and Bower).

Because MSE' and the mean square for the A, or trial, effect (MSA) are not independent (Shine, 1975) due to MSE' being based on successive trial means, MSE' cannot be used to test for the main effect due to trials. The test proposed by Shine and Bower (1971) for the trial main effect is the mean square successive difference (MSSD) test due to Bennet and Franklin (1961):

$$\eta = \frac{\text{MSSD}}{MS_A} = \frac{\dfrac{1}{I-1} \sum_{i=1}^{I-1} (\bar{y}_{i+1} - \bar{y}_i)^2}{\dfrac{1}{I-1} \sum_{i=1}^{I} (\bar{y}_i - \bar{y})^2}.$$

Critical values of η for 5 percent and 1 percent for one-tailed tests are given in Bennet and Franklin (1961), and, for $I > 25$,

$$z = (1 - \eta/2)\sqrt{(I - 1)(I + 1)/(I - 2)}$$

will be approximately normally distributed. A two-tailed MSSD test is supposed to be a test for the main effect of trials and an upper tailed MSSD test is supposed to be a test of the slow change assumption. Modifications to these tests have been proposed by Shine (1974, 1976, 1977), and extension has been made to higher order ANOVA models (Shine, 1973). A note of caution, however, is that Bennet and Franklin (1961) originally present the MSSD test using successive differences of observations, not means, and they present it as a test to detect nonrandomness. Thus, the MSSD should be sensitive to nonzero serial correlation.

Several authors have criticized the independence assumption common to the models presented by Gentile et al. (1972) and Shine and Bower (1971). Kratochwill et al. (1974), Hartmann (1974), Thoreson and Elashoff (1974), and Keselman and Leventhal (1974) all focused attention on the independence assumption and problems associated with use of ANOVA-based models in the presence of nonindependence. Some presented modifications or alternatives but none of these authors showed any empirical or comparative research on these ANOVA-based models. Hartmann (1974) states that (1) researchers should use only the one-way ANOVA model and F test if the assumption of independence is met, "until such time as the nature and extent of the violations of the F test are more fully examined," and (2) if a researcher still wants to use ANOVA models when independence assumptions are not met, "then he should use either the relatively unexplored but more sophisticated ANOVA model suggested by Shine and Bower (1971)" or a variation that uses preliminary testing for lack of independence.

Hartmann's (1974) modification includes using only asymptotic responses with 12 or more stable data points per phase, preliminary tests for nonzero serial correlation for at least lag-one within each phase and for nonzero cross correlations of at least lag zero and one of two different ANOVA designs on the stable data. The simple preliminary testing procedure presented earlier uses only tests on within-phase, lag-one serial correlation and used all the data in the ANOVA F test. Thus, one proceeds with the ANOVA F test only if the preliminary test(s) are nonsignificant (indicating compliance with the independence assumption). Of course, one difficulty with such modifications is the question of what to do if the preliminary test(s) are significant. Does one proceed with caution or switch to another procedure? Another question to be raised is the influence on such tests of nonsignificant serial correlation. If the correlation is so low that it is not detected by the preliminary tests, is it not also problematic for the main effect tests of interest?

The preliminary testing procedure is offered as acceptable statistical analysis by Kazdin (1982) and is intimated as appropriate by Hersen and Barlow (1976). Hartmann's (1974) suggestion of the more complicated preliminary testing method was accompanied by the indication that the Shine-Bower model could be used as an alternative. Gottman and Glass (1978) indicate that the preliminary testing suggestions offered by Hartmann (1974) would be appropriate for certain time series, those that have lag one as the only large serial correlation. Shine and Bower (1971) indicate that "any such correlation can be carried, in a manner similar to that of the standard repeated measures design, by certain effects in the proposed design" (p. 107). Kazdin (1982, pp. 319–321) labels the Shine-Bower model as an alternative or option that uses F to deal with the problem of serial dependency and that is more complex than the preliminary testing procedure. Whether or not Kazdin (1982) suggests use of the Shine-Bower model is questionable. Other authors either uncritically list the Shine-Bower procedure along with other statistical methods for $N = 1$ data (Edgington, 1980; Kratochwill, 1978) or clearly deny the appropriateness of *any* ANOVA-based method for $N = 1$ data (Levin, Marascuilo, & Hubert, 1978), including the Shine-Bower model (also see Gottman & Glass, 1978). Negative comments on these procedures are usually based on the ANOVA assumption of independent errors and the likelihood of violating this assumption with most $N = 1$ research. None of these sources offer results from empirical or analytical research on the preliminary testing procedure or the Shine-Bower model, and are not in accord on the worthiness (or worthlessness) of these models. Given the ambivalence of the literature on these methods, the present research has as its purpose answering questions about robustness of the Shine-Bower tests, ANOVA F test, and the simple preliminary testing procedure. The ANOVA F test is included as a standard to compare to the known results (Box, 1954). These tests will be examined for various patterns/levels of serial correlation, various patterns of phase and trial means, normal and exponential distributions, and equal and unequal phase variances.

Procedure

Monte Carlo simulation was used to study the Shine-Bower F_B, F_{AB}, MSSD (two tailed), MSSD (one tailed), the ANOVA F, and a simple preliminary testing procedure. The simple preliminary testing procedure consisted of separate preliminary tests using the procedure due to Bartlett (1946) (see Kendall & Stuart, 1966, p. 432), on each of the within-phase lag-one serial correlations and proceeding to $F_B = MS_B/MS_W$ only if all preliminary tests were nonsignificant. Only within-phase correlation is examined due to the influence of intervention effects on the serial correlations. If any of the preliminary tests were significant, F_B was not computed, and essentially no decision was made. For all statistics, two different design sizes were used. The smaller design had $J = 4$ levels of phases with $I = 10$ trials per phase, and the larger design had $J = 4$ levels of phases with $I = 30$ trials per phase. Generally, the sampling distributions of the statistics were simulated using computer generated data from a pseudo-random number generator. Random unit-interval uniform variates (Chen, 1971) were generated and then transformed into random variates with mean zero and variance one (z) such that z was distributed as a unit normal (Box & Muller, 1958) or the exponential (Lehmann & Bailey, 1968). These z were then given the desired variance-covariance matrix (**C**) transformed from a simplex-patterned correlation matrix (Guttman, 1955). Because data from a single-subject are characterized by decreasing correlation the further apart the positions of the scores, it seemed that a matrix fitting Guttman's simplex would be appropriate as a model for a correlation matrix for $n = 1$ data. Using a gram-factor decomposition of **C**, then

$$\mathbf{C} = \mathbf{FF}' = \mathbf{QDQ}' = (\mathbf{QD}^{1/2})(\mathbf{D}^{1/2}\mathbf{Q}'), \qquad (1)$$

where **Q** is a matrix of eigenvectors of **C**, and **D** is a symmetric diagonal matrix of the eigenvalues of **C**. If we let $\mathbf{F} = \mathbf{QD}^{1/2}$ and **Z** be the vector of variates described above, then $\mathbf{G} = \mathbf{ZF}'$ has variance-covariance matrix **C**, as given by

$$E(\mathbf{G}'\mathbf{G}) = E(\mathbf{FZ}'\mathbf{ZF}') = \mathbf{F}E(\mathbf{Z}'\mathbf{Z})\mathbf{F}' = \mathbf{FIF}' = \mathbf{FF}' = \mathbf{C}. \qquad (2)$$

Here E is the expected-value operator. Thus, **G** has the desired property of being distributed as multivariate normal or multivariate exponential with variance-covariance matrix **C**. The matrix **G** has dimensions of number-of-subjects by number-of-variables, where number-of-subjects is set equal to one for single-subject data and number-of-variables is equal to the total number of trials, 40 or 120.

By specifying **C**, both the desired simplex correlation pattern and the desired variances were specified. For the four phases, the equal variances were arbitrarily chosen to be $\sigma^2 = 15$. The unequal variances were 3, 14, 16, and 27 for σ_j^2, $j = 1$ to 4, respectively. Variances for each trial were constant within a phase.

The patterns of serial correlation were chosen to give a zero correlation pattern and three nonzero patterns which had the simplex form. Examples of these are given in Table I. The three nonzero patterns were selected to represent not only increasing lag-one serial correlations, but also differing number of large higher order lag serial correlations. The low pattern had no serial correlations larger than .3044, the .05 two-tailed critical value of r, $df = 40$. The other patterns had various numbers of large serial correlations (see Table I).

Values of the means for particular combinations of phases and trials were manipulated by adding values of μ_{ij} to **G**. For the null hypothesis case of equal phase means, all trials in all phases had constant $\mu_{ij} = 0$. Two non-null cases were examined: an ABAB pattern and a linear pattern. For the pattern of phase means given by ABAB, all trials in a given phase had the same mean. The phase means were 1.6, −1.6, 1.6, and −1.6, respectively. The linear pattern of phase means was −2.25, −.75, .75, and 2.25, with all trials in a given phase having the same mean. For the learning curve case, the *trial* means within a phase represented a gradual increase typical of a learning curve (see Table II). All phases had the same pattern of trial means, giving another example of the null hypothesis. The last pattern of means examined was computed from the learning curve means such that the assumption of slow change was met ($\mu_{ij} = \mu_{i+1,j}$ for odd i). These slow-change means also represented the null hypothesis (see Table II).

For each combination of design, serial correlation pattern, variances, and distribution, the statistics were simulated by running 1000 replications of a pseudoexperiment. For each replication, a vector **G** of IJ scores was generated as if it had come from a single subject, the statistics were computed and compared to critical values, and rejections were counted. The proportion of rejections in 1000 replications is an estimate of the probability of either a Type I error or a correct rejction (power). All mean patterns were investigated for the same set of generated data to reduce variability in the results for different mean patterns. That is, for **G** the data were formed by **G** + μ_{ij}, and all mean patterns were based on the same raw data **G**. Each new combination of the other variables resulted in a new **G**.

For each replication, the serial correlations for all the *IJ* scores for lag 1-20 were computed for the vector **G**. These serial correlations were then averaged over the 1000 replications to give estimates of the actual serial correlations for the generated data. Example of these values for lag 1-10 were given in Table I.

Results

Results in the form of proportions as estimates of probabilities (empirical probabilities) are given in Tables III–X. For the mean patterns of equal, learning curve, and slow change, these empirical probabilities are estimates of the probability of a Type I error, or α, for all tests except the MSSD upper tailed test. The learning-curve data are a violation of the slow-change assumption, and the proportions given for the MSSD upper tailed test are estimates of power. For the ABAB and linear patterns of means, the empirical probabilities are estimates of power for the Shine-Bower F_B, ANOVA F and preliminary-testing F and estimates of α for the other tests. For zero serial correlations, all tests give reasonable control of α. The α values for the unequal variances normal distribution case were slightly inflated for the ANOVA and preliminary-testing tests, and the α values for the exponential distribution (equal variances) were slightly conservative for all tests. Combinations of unequal variances and the exponential distribution showed that the effect of unequal variances was more potent and yielded slightly liberal α values for ANOVA and preliminary-testing tests. For zero serial correlation, these tests were more powerful than the Shine-Bower F_B.

For nonzerro serial correlation patterns, all tests showed distinct sensitivity to increasing degree of correlation. All tests except the MSSD upper tailed test were excessively liberal for nonzero serial correlation. For example, the Shine-Bower F_B gave α values of .497, .768, and .995 for low, medium, and high serial correlation, respectively, for equal means, normal distribution, and equal variances for $J = 4$, $I = 10$. While the proportion of rejections out of 1000 replications for the preliminary-testing F decreased from .591 for low correlation to .129 for high correlation, the effective rejection rate of proportion of rejections out of the replications for which the F test was actually computed (those having all within-groups, lag-one serial correlation nonsignificant) increased from .633 for low correlation to .985 for high correlation. For medium or high correlation, the preliminary testing procedure usually would not progress to doing the F test, and if progress to the F test was allowed, a high percentage of false rejections would result. The learning-curve and slow-change

TABLE I
Serial Correlation Patterns

J = 4, I = 10

Lag	Zero	Low	Medium	High
1	-.0222	.2812	.6953	.8659
2	-.0270	.2446	.4110	.7428
3	-.0379	.2075	.1379	.6340
4	-.0267	.1767	-.1441	.5318
5	-.0189	.1518	-.1159	.4366
6	-.0330	.1327	-.0994	.3467
7	-.0159	.0961	-.0919	.2634
8	-.0360	.0713	-.0869	.1910
9	-.0287	.0484	-.0828	.1245
10	-.0129	.0307	-.0793	.0647

J = 4, I = 30

Lag	Zero	Medium	High
1	.0579	.4243	.7358
2	.0535	.3736	.4758
3	.0530	.3331	.2186
4	.0526	.2909	-.0398
5	.0416	.2455	-.0313
6	.0423	.2009	-.0247
7	.0340	.1545	-.0225
8	.0345	.1111	-.0218
9	.0308	.0725	-.0237
10	.0287	.0291	-.0245

TABLE II

Trial Means for I Trials for Learning-Curve and Slow-Change Conditions. All levels of the Experimental Factor (Phases) had the same pattern of Trial Means.

I = 10			I = 30	
Learning-Curve	Slow-Change	Trial	Learning-Curve	Slow-Change
-.22	-.18	1	-.36	-.335
-.14	-.18	2	-.31	-.335
-.08	-.055	3	-.26	-.240
-.03	-.055	4	-.22	-.240
.03	.045	5	-.19	-.180
.06	.045	6	-.17	-.180
.08	.09	7	-.13	-.115
.10	.09	8	-.10	-.115
.10	.10	9	-.08	-.075
.10	.10	10	-.07	-.075
		11	-.04	-.025
		12	-.01	-.025
		13	.02	.03
		14	.04	.03
		15	.06	.065
		16	.07	.065
		17	.08	.085
		18	.09	.085
		19	.10	.105
		20	.11	.105
		21	.12	.125
		22	.13	.125
		23	.14	.14
		24	.14	.14
		25	.14	.14
		26	.14	.14
		27	.14	.14
		28	.14	.14
		29	.14	.14
		30	.14	.14

TABLE III

Empirical Probabilities for 1000 Replications for the Shine-Bower $F_B(B)$, Shine-Bower $F_{AB}(AB)$, ANOVA $F(F)$, Preliminary Testing $F(PT)$, MSSD two-tailed (N), and MSSD upper-tailed (U) Tests, Four Phases by 10 Trials with a Normal Distribution and Equal Phase Variances.

		Correlation Patterns			
Means	Tests	Zero	Low	Medium	High
Equal	B	.049	.497	.768	.995
	AB	.044	.075	.362	.805
	F [a]	.047	.636	.608	.966
	PT [b]	.047 (988)	.591 (934)	.163 (211)	.129 (131)
	N	.097	.125	.687	.725
	U	.047	.036	.002	.002
ABAB	B	.311	.614	.867	.998
	AB	.044	.075	.362	.805
	F	.525	.770	.760	.985
	PT [a][b]	.519 (988)	.717 (934)	.182 (211)	.129 (131)
	N	.097	.125	.687	.725
	U	.047	.036	.002	.002
Linear	B	.318	.544	.843	.993
	AB	.044	.075	.362	.805
	F	.546	.691	.737	.965
	PT [a][b]	.540 (988)	.643 (934)	.179 (211)	.126 (131)
	N	.097	.125	.687	.725
	U	.047	.036	.002	.002
Learning Curve	B	.049	.496	.768	.995
	AB	.047	.074	.361	.795
	F [a]	.047	.635	.605	.968
	PT[b]	.047 (988)	.589 (930)	.162 (201)	.135 (136)
	N	.096	.133	.680	.733
	U	.050	.038	.002	.001
Slow Change	B	.049	.497	.768	.995
	AB	.044	.075	.362	.805
	F	.047	.635	.605	.968
	PT[a][b]	.047 (989)	.589 (929)	.160 (201)	.137 (138)
	N	.096	.135	.680	.731
	U	.049	.038	.002	.001

[a] Values in parentheses are the number of replications with all within-phase lag 1 serial correlations nonsignificant.

[b] Nominal probability = .10, all others, .05.

TABLE IV

Empirical Probabilities for 1000 *Replications for the Shine-Bower $F_B(B)$, Shine-Bower $F_{AB}(AB)$, ANOVA $F(F)$, Preliminary Testing $F(PT)$, MSSD two-tailed (N), and MSSD upper-tailed (U) Tests, Four Phases by* 10 *Trials with a Normal Distribution and Unequal Phase Variances*

		Correlation Patterns			
Means	Tests	Zero	Low	Medium	High
	B	.053	.506	.751	.986
	AB	.049	.075	.352	.738
Equal	F	.063	.611	.611	.962
	PT[a]	.063 (988)	.578 (946)	.153 (217)	.121 (126)
	N[b]	.097	.126	.661	.712
	U	.048	.041	.000	.001
	B	.306	.618	.857	.994
	AB	.049	.075	.352	.738
ABAB	F	.506	.744	.764	.978
	PT[a]	.500 (988)	.705 (946)	.185 (217)	.124 (126)
	N[b]	.097	.126	.661	.712
	U	.048	.041	.000	.001
	B	.333	.578	.836	.991
	AB	.049	.075	.352	.738
Linear	F	.552	.687	.714	.973
	PT[a]	.548 (988)	.648 (946)	.174 (217)	.118 (126)
	N[b]	.097	.126	.661	.712
	U	.048	.041	.000	.001
	B	.056	.505	.747	.986
	AB	.049	.072	.351	.734
Learning Curve	F	.063	.604	.609	.960
	PT[a]	.063 (986)	.570 (945)	.147 (218)	.116 (122)
	N[b]	.100	.134	.676	.718
	U	.048	.041	.001	.000
	B	.053	.506	.751	.986
	AB	.049	.075	.352	.738
Slow Change	F	.062	.605	.609	.960
	PT[a]	.062 (986)	.572 (945)	.145 (213)	.115 (121)
	N[b]	.103	.136	.678	.719
	U	.048	.040	.001	.001

[a] Values in parentheses are the number of replications with all within-phase lag 1 serial correlations nonsignificant.

[b] Nominal probability = .10, all others, .05.

TABLE V

Empirical Probabilities for 1000 *Replications for the Shine-Bower $F_B(B)$, Shine-Bower $F_{AB}(AB)$, ANOVA $F(F)$, Preliminary Testing $F(PT)$, MSSD two-tailed (N), and MSSD upper-tailed (U) Tests, Four Phases by* 10 *Trials with an Exponential Distribution and Equal Phase Variances*

		Correlation Patterns			
Means	Tests	Zero	Low	Medium	High
	B	.038	.451	.685	.992
	AB	.043	.117	.377	.735
Equal	F	.032	.557	.517	.964
	PT[a]	.032 (979)	.505 (919)	.154 (267)	.167 (174)
	N [b]	.089	.193	.647	.711
	U	.046	.063	.003	.002
	B	.312	.640	.796	.997
	AB	.043	.117	.377	.735
ABAB	F	.565	.792	.697	.976
	PT[a]	.554 (979)	.734 (919)	.196 (267)	.171 (174)
	N [b]	.089	.193	.647	.711
	U	.046	.063	.003	.002
	B	.334	.448	.832	.988
	AB	.043	.117	.377	.735
Linear	F	.571	.503	.730	.929
	PT[a]	.559 (979)	.449 (919)	.221 (267)	.163 (174)
	N [b]	.089	.193	.647	.711
	U	.046	.063	.003	.002
	B	.039	.451	.682	.992
	AB	.040	.117	.380	.730
Learning	F	.031	.557	.520	.960
Curve	PT[a]	.031 (978)	.504 (920)	.145 (257)	.178 (185)
	N [b]	.099	.210	.672	.690
	U	.051	.066	.004	.001
	B	.038	.451	.685	.992
	AB	.043	.117	.377	.735
Slow	F	.031	.557	.520	.961
Change	PT[a]	.031 (978)	.505 (922)	.144 (256)	.173 (180)
	N [b]	.100	.206	.666	.693
	U	.051	.065	.004	.002

[a] Values in parentheses are the number of replications with all within-phase lag 1 serial correlations nonsignificant.

[b] Nominal probability = .10; all others, .05.

Table VI
Empirical Probabilities for 1000 *Replications for the Shine-Bower* $F_B(B)$, *Shine-Bower* $F_{AB}(AB)$, *ANOVA* $F(F)$, *Preliminary Testing* $F(PT)$, *MSSD two-tailed* (N), *and MSSD upper-tailed* (U) *Tests, Four Phases by* 10 *Trials with an Exponential Distribution and Unequal Phase Variances*

		Correlation Patterns			
Means	Tests	Zero	Low	Medium	High
	B	.042	.451	.674	.992
	AB	.038	.076	.374	.691
Equal	F	.060	.595	.501	.960
	PT[a]	.060 (979)	.516 (882)	.122 (222)	.135 (146)
	N [b]	.085	.176	.635	.693
	U	.045	.066	.008	.002
	B	.353	.632	.858	.998
	AB	.038	.076	.347	.691
ABAB	F	.546	.803	.787	.995
	PT[a]	.531 (979)	.709 (882)	.188 (222)	.145 (146)
	N [b]	.085	.176	.635	.693
	U	.045	.066	.008	.002
	B	.320	.436	.773	.986
	AB	.038	.076	.347	.691
Linear	F	.591	.542	.661	.955
	PT[a]	.581 (979)	.471 (882)	.156 (222)	.140 (146)
	N [b]	.085	.176	.635	.693
	U	.045	.066	.008	.002
	B	.042	.443	.675	.991
	AB	.038	.074	.342	.696
Learning	F	.062	.595	.498	.959
Curve	PT[a]	.062 (976)	.518 (882)	.129 (234)	.135 (143)
	N [b]	.082	.187	.644	.686
	U	.042	.065	.003	.004
	B	.042	.451	.674	.992
	AB	.038	.076	.347	.691
Slow	F	.062	.595	.498	.959
Change	PT[a]	.062 (977)	.515 (879)	.127 (232)	.139 (149)
	N [b]	.081	.191	.643	.677
	U	.042	.066	.003	.003

[a] Values in parentheses are the number of replications with all within-phase lag 1 serial correlations nonsignificant.

[b] Nominal probability = .10; all others, .05.

TABLE VII

Empirical Probabilities for 1000 Replications for the Shine-Bower $F_B(B)$, Shine-Bower $F_{AB}(AB)$, ANOVA $F(F)$, Preliminary Testing $F(PT)$, MSSD two-tailed (N), and MSSD upper-tailed (U) Tests, Four Phases by 30 Trials with a Normal Distribution and Equal Phase Variances

			Correlation Patterns	
Means	Tests	Zero	Low	Medium
Equal	B	.050	.815	.893
	AB	.046	.248	.875
	F	.042	.735	.631
	PT[a]	.042 (965)	.138 (178)	.000 (000)
	N [b]	.091	.511	.990
	U	.038	.006	.000
ABAB	B	.550	.845	.929
	AB	.046	.248	.875
	F	.648	.781	.722
	PT[a]	.628 (965)	.148 (178)	.000 (000)
	N [b]	.091	.511	.990
	U	.038	.006	.000
Linear	B	.439	.834	.938
	AB	.046	.248	.875
	F	.507	.776	.735
	PT[a]	.490 (965)	.149 (178)	.000 (000)
	N [b]	.091	.511	.990
	U	.038	.006	.000
Learning Curve	B	.053	.836	.873
	AB	.056	.276	.872
	F	.047	.769	.583
	PT[a]	.046 (970)	.132 (170)	.000 (000)
	N [b]	.114	.502	.995
	U	.059	.003	.000
Slow Change	B	.053	.834	.873
	AB	.058	.278	.871
	F	.047	.769	.583
	PT[a]	.046 (970)	.132 (171)	.000 (000)
	N [b]	.115	.503	.995
	U	.060	.003	.000

[a] Values in parentheses are the number of replications with all within-phase lag 1 serial correlations nonsignificant.

[b] Nominal probability = .10; all others, .05.

Table VIII

Empirical Probabilities for 1000 Replications for the Shine-Bower $F_B(B)$, Shine-Bower $F_{AB}(AB)$, ANOVA $F(F)$, Preliminary Testing $F(PT)$, MSSD two-tailed (N), and MSSD upper-tailed (U) Tests, Four Phases by 30 Trials with a Normal Distribution and Unequal Phase Variances

		Correlation Patterns		
Means	Tests	Zero	Low	Medium
Equal	B	.060	.797	.878
	AB	.043	.243	.880
	F	.056	.730	.557
	PT[a]	.056 (965)	.137 (190)	.000 (000)
	N[b]	.098	.539	.992
	U	.049	.003	.000
ABAB	B	.500	.841	.908
	AB	.043	.243	.880
	F	.602	.789	.680
	PT[a]	.577 (965)	.153 (190)	.000 (000)
	N[b]	.098	.539	.992
	U	.049	.003	.000
Linear	B	.439	.841	.937
	AB	.043	.243	.880
	F	.512	.793	.735
	PT[a]	.496 (965)	.154 (190)	.000 (000)
	N[b]	.098	.539	.992
	U	.049	.003	.000
Learning Curve	B	.053	.779	.870
	AB	.045	.259	.853
	F	.055	.689	.575
	PT[a]	.054 (967)	.115 (165)	.000 (000)
	N[b]	.098	.529	.990
	U	.051	.006	.000
Slow Change	B	.052	.780	.871
	AB	.045	.262	.853
	F	.055	.689	.575
	PT[a]	.054 (967)	.115 (165)	.000 (000)
	N[b]	.100	.528	.990
	U	.052	.006	.000

[a] Values in parentheses are the number of replications with all within-phase lag 1 serial correlations nonsignificant.

[b] Nominal probability = .10, all others, .05.

Table IX

Empirical Probabilities for 1000 Replications for the Shine-Bower $F_B(B)$, Shine-Bower $F_{AB}(AB)$, ANOVA $F(F)$, Preliminary Testing $F(PT)$, MSSD two-tailed (N), and MSSD upper-tailed (U) Tests, Four Phases by 30 Trials with an Exponential Distribution and Equal Phase Variances

		Correlation Patterns		
Means	Tests	Zero	Low	Medium
	A	.045	.771	.829
	AB	.052	.307	.818
Equal	F	.041	.675	.511
	PT[a]	.040 (950)	.194 (282)	.000 (000)
	N[b]	.090	.462	.970
	U	.044	.019	.000
	A	.561	.866	.937
	AB	.052	.307	.818
ABAB	F	.641	.829	.739
	PT[a]	.612 (950)	.238 (282)	.000 (000)
	N[b]	.090	.462	.970
	U	.044	.019	.000
	B	.481	.714	.933
	AB	.052	.307	.818
Linear	F	.542	.630	.709
	PT[a]	.513 (950)	.160 (282)	.000 (000)
	N[b]	.090	.462	.970
	U	.044	.019	.000
	B	.052	.747	.843
	AB	.043	.305	.791
Learning	F	.039	.662	.490
Curve	PT[a]	.037 (951)	.201 (286)	.000 (000)
	N[b]	.090	.476	.972
	U	.042	.020	.000
	B	.052	.747	.843
	AB	.044	.306	.792
Slow	F	.039	.662	.490
Change	PT[a]	.037 (952)	.202 (287)	.000 (000)
	N[b]	.091	.475	.971
	U	.042	.019	.000

[a] Values in parentheses are the number of replications with all within-phase lag 1 serial correlations nonsignificant.

[b] Nominal probability = .10; all others, .05.

TABLE X

Empirical Probabilities for 1000 Replications for the Shine-Bower $F_B(B)$, Shine-Bower $F_{AB}(AB)$, ANOVA $F(F)$, Preliminary Testing $F(PT)$, MSSD two-tailed (N), and MSSD upper-tailed (U) Tests, Four Phases by 30 Trials with an Exponential Distribution and Unequal Phase Variances

		Correlation Patterns		
Means	Tests	Zero	Low	Medium
	B	.068	.721	.814
	AB	.037	.279	.836
Equal	F	.053	.666	.467
	PT[a]	.051 (950)	.117 (182)	.000 (000)
	N[b]	.086	.486	.976
	U	.044	.018	.000
	B	.526	.821	.862
	AB	.037	.279	.836
ABAB	F	.619	.773	.555
	PT[a]	.589 (950)	.147 (182)	.000 (000)
	N[b]	.086	.486	.976
	U	.044	.018	.000
	B	.422	.754	.959
	AB	.037	.279	.836
Linear	F	.507	.696	.790
	PT[a]	.482 (950)	.127 (182)	.000 (000)
	N[b]	.086	.486	.976
	U	.044	.018	.000
	B	.060	.690	.810
	AB	.033	.256	.837
Learning Curve	F	.055	.619	.478
	PT[a]	.054 (946)	.120 (182)	.001 (001)
	N[b]	.092	.496	.971
	U	.038	.010	.000
	B	.060	.691	.812
	AB	.033	.260	.837
Slow Change	F	.055	.619	.478
	PT[a]	.054 (946)	.121 (183)	.001 (001)
	N[b]	.089	.495	.972
	U	.038	.010	.000

[a] Values in parentheses are the number of replications with all within-phase lag 1 serial correlations nonsignificant.

[b] Nominal probability = .10; all others, .05.

mean patterns gave similar results. These same results were replicated for unequal variances, exponential distribution, and all cases for the larger design. Given the excessive α values in the face of nonzero serial correlation, the value of a discussion of power is questionable. The MSSD upper tailed test was increasingly conservative as a function of increasing serial correlation and was not sensitive to violation of the slow-change assumption.

Conclusions

Considerable attention has been given to the problem of data analysis for $N = 1$ designs. Several authors mentioned or alluded to the need for further research on traditional ANOVA-based tests, and Hartmann (1974) suggested using the Shine-Bower tests when independence assumptions are not met. Shine and Bower (1971) indicated that serial correlation in the data would be carried "by certain effects" without giving any methodology for separating the effects of nonzero serial correlation from the effects of the experimental factor or interaction. Subsequent articles by Shine (1980, 1981, 1982) argue for two single-subject behavior functions, one of which is an independent error model such as claimed for the Shine-Bower model. This paper attempts to clarify that use of the Shine-Bower analysis should be restricted only to data that fits the independent error situation. The present research shows that the Shine-Bower, ANOVA, and preliminary testing tests for the experimental factor are seriously influenced by violation of the independence assumption. Positive nonzero serial correlation causes excessively liberal α values, and these tests are not robust to even nonsignificant serial correlation, which usually would not be detected by tests for serial correlation. Thus, the results of Box (1954) and Hibbs (1974) generalize to these ANOVA-based methods for $n = 1$ data: the statistics are inflated by the presence of nonzero serial correlation. Because none of these procedures can be recommended for hypothesis testing in single-subject research with positive lag-one serial correlation, the researcher with such data may turn to the other methods mentioned earlier. For the researcher who wants to use statistical methods for serially correlated $n = 1$ data, time series should suffice because it is designed to model dependency such as used in the present research. For the researcher who has data that are not serially correlated, the ANOVA F test or the preliminary testing procedure provide more power than the Shine-Bower tests.

Acknowledgment

This research was supported by NIMH grant number R01MH33183.

References

Barlow, D. H., & Hersen, M. Single case experimental designs. *Archives of General Psychiatry*, 1973, *29*, 319–325.

Bartlett, M. S. On the theoretical specification and sampling properties of autocorrelated time-series. *Supplemental Journal of the Royal Statistical Society*, 1946, *8*, 27.

Bennet, C. A., & Franklin, N. L. *Statistical analysis in chemistry and the chemistry industry*. New York: John Wiley, 1961.

Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems. II. Effects of inequality of variance and of correlation between errors in the two-way classification. *Annals of Mathematical Statistics*, 1954, *25*, 484–498.

Box, G. E. P., & Jenkins, G. M. *Time series analysis*: *Forecasting and control* (Rev. ed.). San Francisco: Holden-Day, 1976.

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

Chassan, J. B. *Research designs in clinical psychology and psychiatry*. New York: Appleton-Century-Crofts, 1967.

Chen, E. H. Random normal number generator for 32-bit-word computers. *Journal of the American Statistical Association*, 1971, *66* (334), 400–403.

Cook, T. D., & Campbell, D. T. *Quasi-experimentation*: *Design and analysis issues for field settings*. Chicago: Rand-McNally, 1979.

Deustch, S. J., & Alt, F. B. The effect of Massachusetts' gun control law on gun-related crimes in the city of Boston. *Evaluation Quarterly*, 1977, *1*, 543–568.

Edgington, E. S. $N = 1$ experiments: Hypothesis testing. *The Canadian Psychologist*, 1972, *2*, 121–134.

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

Friesma, H. P., Caporaso, J. A., Lineberry, R. L., & Goldstein, G. S. *Aftermath*: *Community impact of natural disasters*. Beverly Hills, Calif.: Sage Publications, 1978.

Gastwirth, J. L., & Rubin, H. Effect of dependence on the level of some one-sample tests. *Journal of the Americal Statistical Association*, 1971, *66*, 816–820.

Gentile, J. T., Roden, C., & Klein, R. An analysis of variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis*, 1972, *5*, 193–198.

Glass, G. V, Tiao, G. C., & Maguire, T. O. Analysis of data on the 1900 revision of German divorce laws as a time-series quasi-experiment. *Law and Society Review*, 1971, *4*, 539–562.

Glass, G. V, Willson, V. L., & Gottman, J. M. *Design and analysis of time-series experiments*. Boulder, Colo.: Colorado Associated University Press, 1975.

Gottman, J. M., & Glass, G. V. Analysis of interrupted time-series experiments. In T. R. Kratochwill (Ed.), *Single subject research*: *Strategies for evaluating change*. New York: Academic Press, 1978.

Gottman, J. M., & McFall, R. M. Self-monitoring effects in a program for potential high school dropouts: A time-series analysis. *Journal of Consulting and Clinical Psychology*, 1972, *39*, 273–281.

Gottman, J. M., & Notarius, C. Sequential analysis of observational data using Markov chains. In T. R. Kratochwill (Ed.), *Single subject research*: *Strategies for evaluating change*. New York: Academic Press, 1978.

Guttman, L. A. A generalized simplex for factor analysis. *Psychometrika*, 1955, *20*, 173–192.

Hartmann, D. P. Forcing square pegs into round holds: Some comments on "An analysis of variance model for the intrasubject replication design." *Journal of Applied Behavior Analysis*, 1974, *7*, 635–638.

Hartmann, D. P., Gottman, J. M., Jones R. R., Gardner, W., Kazdin, A. E., & Vaught, R. S. Interrupted time-series analysis and its application to behavioral data. *Journal of Applied Behavior Analysis*, 1980, *13*, 543–559.

Hedberg, A. G., Walker, C. E., & Ehrman, J. Annotated bibliography of statistics for clinicians. *Profession Psychology*, 1975, *6*(1), 96–100.

Hersen, M., & Barlow, D. H. *Single case experimental designs: Strategies for studying behavior change*. New York: Pergamon, 1976.

Hibbs, D. A., Jr. Problems of statistical estimation and causal inference on time-series regression models. In H. L. Costner (Ed.), *Sociological Methodology-1974*. San Francisco: Jossey-Bass, 1974.

Huitema, B. E. Autocorrelation in behavior modification data: Wherefore are thou? *Behavioral Assessment*, 1983.

Jones, R. R., Vaught, R. S., & Weinrott, M. R. Time-series analysis in operant research. *Journal of Applied Behavior Analysis*, 1977, *10*, 151–166.

Juliano, D. B., & Gentile, J. T. Will the real hyperactive child please sit down? Problems of diagnosis and remediation. *Child Study Journal Monograph*, 1974 (*m*1–6), 1–38.

Kazdin, A. E. *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press, 1982.

Kendall, M. G., & Stuart, A. *The advanced theory of statistics* (Vol. 3). London: Charles Griffin & Company, 1966.

Keselman, J., & Leventhal, L. Concerning the statistical procedures enumerated by Gentile et al.: Another perspective. *Journal of Applied Behavior Analysis*, 1974, *7*(4), 643–645.

Kratochwill, T. R. $N = 1$: An alternative research strategy for school psychologists. *Journal of School Psychology*, 1977, *15*, 239–249.

Kratochwill, T. R. Foundations of time-series research. In T. R. Kratochwill (Ed.), *Single subject research: Strategies for evaluating change*. New York: Academic Press, 1978.

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

Lehmann, R. S., & Bailey, D. E. Digital computing: *Fortran IV and its applications in behavioral science*. New York: John Wiley and Sons, 1968.

Leitenberg, H., Agras, W. S., & Thomson, L. A sequential analysis of the effect of selective positive reinforcement in modifying anorexia nervosa. *Behavior Research and Therapy*, 1968, *6*, 211–218.

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

Parsonson, B. S., & Baer, D. M. The analysis and presentation of graphic data. In T. R. Kratochwill (Ed.), *Single subject research: Strategies for evaluating change*. New York: Academic Press, 1978.

Shine, L. C. A multi-way analysis of variance for single-subject designs. *Educational and Psychological Measurement*, 1973, *33*, 633–636.

Shine, L. C. An extension of the Shine combined analysis variance. *Educational and Psychological Measurement*, 1974, *34*, 47–52.

Shine, L. C. Independence problems for certain tests based on the Shine-Bower error term. *Educational and Psychological Measurement*, 1975, *35*, 535–537.

Shine, L. C. An alternative statistic for testing the slow change assumption of the Shine-Bower error term. *Educational and Psychological Measurement*, 1976, *34*, 611–614.

Shine, L. C. Removing the restriction of an even number of trials for the Shine-Bower single-subject ANOVA. *Educational and Psychological Measurement*, 1977, *37*, 873–875.

Shine, L. C. On two fundamental single-subject behavior functions. *Educational and Psychological Measurement*, 1980, *40*, 63–72.

Shine, L. C. Integrating the study of Shine's actualized and pure single-subject behavior functions. *Educational and Psychological Measurement*, 1981, *41*, 673–685.

Shine, L. C. An illustration of how the effects of serial dependencies are handled in analyses of Shine's pure and actualized single-subject behavior functions. *Educational and Psychological Measurement*, 1982, *42*, 87–94.

Shine, L. C., & Bower, S. M. A one-way analysis of variance for single-subject designs. *Educational and Psychological Measurement*, 1971, *31*, 105–113.

Shine, L. C., Wiant, J., & Dapalito, F. Effect of learning on hemispheric dominance and free recall in a single subject. *Psychological Reports*, 1972, *31*, 227–230.

Sterman, M. B. Neurophysiologic and clinical studies of sensorimotor EEG biofeedback training: Some effects of epilepsy. In L. Birk (Ed.), *Biofeedback: Behavioral medicine*. New York: Grune & Stratton, 1973.

Thoreson, C. E., & Elashoff, J. D. An analysis of variance model for the intrasubject replication design: Some additional comments. *Journal of Applied Behavior Analysis*, 1974, 7, 639–641.

Authors

TOOTHAKER, LARRY E. Professor, Department of Psychology, University of Oklahoma, Norman, OK 73019. *Specializations*: Randomization tests, individual comparisons, robustness.

BANZ, MARTHA L. Instructor, Bethany Nazarene College, Bethany, OK 73008. *Specialization*: Quantitative psychology.

NOBLE, CYNTHIA A. Instructor, University of Science and Arts of Oklahoma, Chickasha, OK 73018. *Specialization*: Developmental psychology.

CAMP, JILL JEWELL. Petroleum Landman, Land Associates, Inc., P.O. Box 2814, Tulsa, OK. *Specializations*: Primate behavior, non-verbal communication.

DAVIS, DIANA. Psychologist, Psychology Department, Warm Springs Rehabilitation Hospital, Box 58, Gonzales, TX 78629. *Specializations*: Neuropsychology, brain injury.