An Empirical Examination of Visual Analysis Procedures for Clinical Practice Evaluation

Jeffrey J. Borckardt
Martin D. Murphy
Michael R. Nash
Darlene Shaw

ABSTRACT. There has been a resurgence of interest in single-subject research designs and analytic tools to help clinicians detect treatment effects. The present study investigates Nugent's (2000) visual analysis procedures, which were designed to aid practitioners in detecting clinical change for the purposes of practice evaluation. The ability of the visual procedures to detect real change in short auto-correlated data streams and the ability of the procedures to help clinicians discern cases when no actual change has occurred were evaluated. Monte Carlo analyses indicate that the power of the visual procedures is acceptable for effect sizes of 2.25 or greater when there are at least 14 data points (7 baseline and 7 treatment) in the data set. The procedures, however, frequently lead to erroneous decisions that effects are present in

Jeffrey J. Borckardt, PhD, is Post-Doctoral Fellow, Medical University of South Carolina, Counseling and Psychological Services (CAPS), 45 Courtenay Drive, Room SS 224, Charleston, SC 29425 (E-mail: borckard@musc.edu). Martin D. Murphy, PhD, is Professor, University of Akron, Dept. of Psychology, Arts and Sciences Bldg., Room 366, Akron, OH 44325 (E-mail: mmurphy@uakron.edu). Michael R. Nash, PhD, is Professor, Psychology Dept., 307 Austin Peay Bldg., the University of Tennessee, Knoxville, TN 37996 (E-mail: mnash@utk.edu). Darlene Shaw, PhD, is Professor & Vice Chair of Education for Dept. of Psychiatry and Behavioral Sciences, Director of Counseling & Psychological Services, and Associate Dean of Student Life, Medical University of South Carolina, CAPS (E-mail: Shawd@musc.edu).

Journal of Social Service Research, Vol. 30(3) 2004
http://www.haworthpress.com/web/JSSR

Digital Object Identifier: 10.1300/J079v30n03_04

data streams when, in fact, there are none. The mean type I error rate across various N's and levels of auto-correlation was .66. As they are currently designed, Nugent's visual analysis procedures make too many type I errors to be useful.

KEYWORDS. Time-series analysis, clinical practice, practice evaluation, visual analysis

INTRODUCTION

There is a growing interest among clinicians in accountability and in demonstrating clinical effectiveness empirically (Schwartz, 1997; Beutler, 2000; Clement, 1994; Callaghan, 2001). As a result, more clinicians and clinical researchers are taking an interest in the single-subject research designs and the unique information they offer (Hilliard, 1993; Marten & Heimberg, 1995; Goldfried & Wolfe, 1998; Lowman, 2001; Gedo, 1999; Howard, 1993). Unfortunately, good (useful and objective) analytic approaches do not seem to be available to clinicians for the evaluation of single-subject time series data (Robey, Shultz, Crawford & Sinner, 1999; Crosbie, 1993, 1994; Bloom, Fischer & Orme, 2003). Most of the available approaches require more data points than is practicable or demonstrate poor type I error control and/or power (McKnight, McKean & Huitema, 2000; Huitema & McKean, 2000; Huitema & McKean, 1991; Robey et al., 1999; Crosbie, 1993, 1994). Additionally, many of the more objective approaches are extremely complicated and require unique statistical tools and know-how to implement (Ostrom, 1990). Subjective approaches, identifying effects as real when they appear compelling, may often work well, but are not guaranteed to give the same results for different observers (Matyas & Greenwood, 1990; Robey et al., 1999; Crosbie, 1993, 1994).

The data sets typically available to practicing clinicians have a number of common characteristics. First, the data streams tend to be extremely short (Huitema & McKean, 1991). Clinicians will usually have very short baseline data streams to work with and, often, limited treatment phase data streams as well. Further, the data points tend to be serially dependent (autocorrelated), resulting in problems with statistical

inference when applying conventional statistics, which assume independence (Franklin, Allison & Gorman, 1996; Robey et al., 1999; Crosbie, 1993, 1994; Ostrom, 1990).

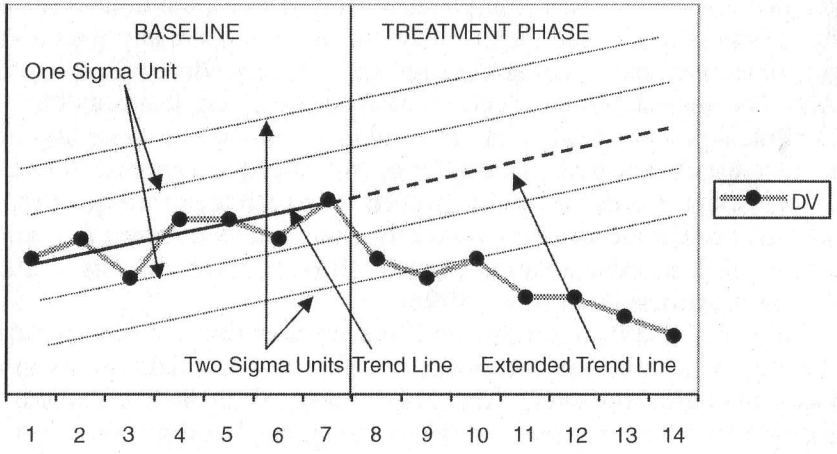
Nugent (2000) addresses the lack of available time series tests, the problems associated with complicated analytic tools, the limited N-sizes in clinical practice, and the problems of serial dependency with a simple, objective, and systematic visual analytic procedure for time series data. The approach is designed for ease-of-use by clinicians and for use in clinical practice evaluation. It requires no special software and little knowledge of statistical matters. Nugent's procedures are broadly-based on statistical process control approaches and are therefore quite representative of the informal methods clinicians and researchers frequently employ in practice evaluation contexts (Bloom, Fischer & Orme, 2003; Franklin, Allison & Gorman, 1996).

Nugent's (2000) approach involves representing baseline trend by drawing a straight line connecting the first and last data points in the baseline phase. However, if the difference between the first and second data points (or the last and second-to-last) in the baseline phase is more than 2.5 times the difference between any other two adjacent data points in the phase, the line is drawn from the mean between points one and two (or drawn to the mean of the last and second-to-last points). Additionally, Nugent (2000) makes accommodations for curvilinear trend. If all (or nearly all) of the data points in the baseline phase fall on one side of the trend line, a line is drawn from the first data point to the lowest (or highest) mid-phase data point. Then, the line is drawn from that mid-phase point to the last point. Once the trend is represented in the baseline phase, the trend line is extended into the treatment phase. See Figure 1 for a labeled illustration of these procedures.

Next, Nugent (2000) proposes a procedure for determining variability in the baseline phase. The user simply finds the baseline data point furthest from the trend line and draws a line parallel to the trend line through that point (the distance between the two lines is called a sigma unit). Another parallel line is drawn on the other side of the trend line the same absolute distance away from the trend line as the first line. Next, two more lines are drawn on each side of the trend line, each one more sigma unit away from the trend line.

Once the trend and variability (sigma units) are identified, the user can evaluate the treatment-phase data points against the baseline properties. Nugent (2000) proposes that four decision rules adapted from statistical process theory be used. The rules are as follows:

FIGURE 1. Labeled Illustration of Nugent's (2000) Visual Analysis Procedure for Detecting Between-Phase Change in Clinical Time Series Data Using a Hypothetical Dependent Variable (DV) Across 14 Weeks.



Rule #1: A significant change is indicated by a data point (in the treatment phase) that falls more than 2 sigma units away from the extended trend line.

Rule #2: Two of three data points falling more than 2 sigma units away from the extended trend line will signify significant change.

Rule #3: Four of five consecutive data points on the same side of the extended trend line, and more than 1 sigma unit away from the extended trend line, signify significant change.

Rule #4: Seven data points in a row that are on the same side of the extended trend line (but not necessarily more than 1 sigma unit away from the extended trend line) are indicative of significant change.

Despite suggesting the application of these decision rules for the determination of between-phase change, Nugent clearly defines the context for the application of his visual procedures:

> The methods described above are best considered in the context of practice evaluation as opposed to that of research. These contexts carry with them different emphases in terms of type I and type II errors, the use of methods, and the types of decisions made and conclusions drawn (Posavac, 1998). *The methods described above should not be used blindly to make decisions about significant change.*

Nonetheless, it seems reasonable and important to investigate the ability of Nugent's visual procedures and the associated decision rules to help clinicians make appropriate decisions regarding the presence or absence of between-phase change. In any context (e.g., laboratory, the consulting room, program evaluation) it is important to know how much confidence one can have in an inference derived from data (i.e., how many times will I be wrong in inferring an effect? How many times will there be a real effect that I miss?). Even if we do not use the conventionally accepted type I and type II error rates (.05 and .20, respectively) as the bench mark to evaluate the procedures against, we expect the procedures to aid in a clinician's ability to be discerning regarding the effectiveness of psychotherapeutic interventions while at the same time offering adequate power to detect real changes when they are, in fact, present. Additionally, we expect the visual procedures to aid clinicians with regard to these domains even when auto-correlation is present in the data.

METHODS AND RESULTS

Monte Carlo analyses are often used to evaluate the properties (usually type I and type II error rates) of statistical procedures. A type I error occurs when a clinician/researcher concludes that a significant effect is present in a data set when there is no real effect. A type II error occurs when one concludes that there is no effect in a data set when, in fact, an effect is present. Typically, Monte Carlo analyses involve the generation of numerous simulated data sets with known properties. These data sets are then evaluated by the statistical approach in question to see how well it performs (e.g., type I or type II error) under a variety of conditions (e.g., when skewness or auto-correlation are present in the data samples). For example, one can generate thousands of completely random data sets (i.e., no effects present) and then investigate how often a statistic or procedure finds a significant effect in that data (type I error).

Monte Carlo analyses were conducted on Nugent's visual procedures with various data stream lengths (total number of data points = 20, 16,

14, 12, 10, and 8) with an equal number of data points in each phase. Additionally, various levels of auto-correlation (ρ) were implemented (−.4, −.2, 0, .2, .4, .8). Auto-correlation occurs in clinical data when un-modeled effects lead errors in successive observations to be correlated. Because positive auto-correlation produces apparent trends in data, its existence has been a longstanding concern (Gottman, 1981; Ostrom, 1990).

The errors for each observation in the simulated streams followed the first order auto-regressive model (Gottman, 1981; McKnight, McKean & Huitema, 2000; Ferron & Sentovich, 2002; Ferron & Onghena, 1996; Crosbie, 1993, 1994) where N = the total number of observations in the data stream taken at time t:

$$e_t = \rho(e_{t-1}) + u_t$$

where t = 2 to N, ρ = programmed level of auto-correlation, e_1 = a normal random deviate N(0,1) generated via the polar method (see Knuth, 1981 for a detailed description of these procedures), and u_t = independent normal random deviate N(0,1) at time t also generated via the polar method (Crosbie, 1993, 1994).

A computer program was designed specifically to carry out Nugent's (2000) visual procedures and apply the four decision rules. Baseline trend was represented as follows: (1) baseline slope was calculated by subtracting the first data point from the last data point in the baseline phase and then dividing by the total number of data points in the phase, (2) if the difference between the first and second data points (or the last and second-to-last) in the baseline phase was more than 2.5 times the difference between any other two adjacent data points in the phase, the mean between the first and second points (and/or the mean of the last and second-to-last points) was used in place of the first (and/or last) data point, and (3) the intercept was equal to the value of the first data point (or mean of the first and second if rule two above was implemented) minus the slope, (4) the baseline data point furthest from the trend line was identified and the distance between it and the trend line defined "one sigma unit," (5) the expected value (i.e., extended trend line) at any point in the data stream was equal to the slope times the data-point number plus the intercept, and (6) data points in the treatment phase were evaluated against the extended trend line plus or minus different numbers of sigma units depending upon the decision rule being evaluated.

Specifically, Nugent's decision rules were evaluated using the following criteria where N = number of data points in the simulated data streams, $n1$ = length of phase-A = $n2$ = length of phase-B, each data point in the simulated data is represented as y_X where x can range from *1 to N*, b = calculated slope of data points in phase-A of the simulated stream, a = calculated intercept of phase-A data, and s = one sigma unit. Each decision rule was only permitted to detect one significant change per data stream.

Rule #1 states that a significant change is indicated by a data point (in the treatment phase) that falls more than two sigma units away from the extended trend line. As such, the program was set up to indicate that Rule #1 detected a significant change under the following condition(s) at any level of x:

$$y_x > a + bx + 2s$$

OR

$$y_x < a + bx - 2s$$

where $x = n1 + 1$ *to* N

Rule #2 states that two of three data points falling more than 2 sigma units away from the extended trend line will signify significant change. Note that Rule #2 overlaps with Rule #1 such that if Rule #2 detects a change, so does Rule #1. The program was set up to indicate that Rule #2 detected a significant change under the following condition(s) for any two out of three consecutive data points:

$$y_x > a + bx + 2s$$

OR

$$y_x < a + bx - 2s$$

where $x = n1 + 1$ *to* N

Rule #3 states that four of five data points on the same side of the extended trend line, and more than 1 sigma unit away from the extended trend line, signify significant change. Nugent (2000) indicates that if the phase n-size is less than 5, all data points in phase-B must fall outside 1sigma unit in order to suggest significant change. When phase n-size was greater than 4, the program was set up to indicate that Rule #3 detected a significant change under the following condition(s) for any four out of five consecutive data points:

$$y_x > a + bx + s$$

$$OR$$

$$y_x < a + bx - s$$

$$\text{where } x = n1 + 1 \text{ to } N$$

When data streams had four data points per phase ($N = 8$ condition), the following had to be true at all levels of x for Rule #3 to detect a significant change:

$$y_x > a + bx + s$$

$$OR$$

$$y_x < a + bx - s$$

$$\text{where } x = n1 + 1 \text{ to } N$$

Rule #4 states that seven data points in a row that are on the same side of the extended trend line (but not necessarily more than 1 sigma unit away from the extended trend line) are indicative of significant change. Nugent (2000) indicates that if phase n-size is less than 7 but all of the phase-B data points are on the same side of the trend line, this pattern suggests change. When N was equal to 20, 16 or 14, the following condition(s) for seven consecutive data points had to be satisfied in order for Rule #4 to indicate change:

$$y_x > a + bx$$

$$OR$$

$$y_x < a + bx$$

$$\text{where } x = n1 + 1 \text{ to } N$$

When N was less than 14, Rule #4 indicated significant change if the following was true at all levels of x:

$$y_x > a + bx$$

$$OR$$

$$y_x < a + bx$$

$$\text{where } x = n1 + 1 \text{ to } N$$

Lastly, if any of the four decision rules indicated change, the Overall detection index indicated change.

Nugent recommends that if all (or nearly all) of the data points in the baseline phase fall on one side of the trend line, that measures be taken to accommodate possible curvilinear trend. This could be problematic when there are very few data points in the baseline phase (e.g., four) as the middle points fall on the same side of the trend line frequently, even when no curvilinear trend exists. Thus, if we attempt to identify curvilinear trend and accommodate it in the fashion recommended by Nugent (2000), we are likely to frequently misrepresent trends. Additionally, no curvilinear trends were programmed into the simulated data streams evaluated in this study and, as such, any identification of curvilinear trend by the visual procedures would be erroneous. So, to be fair to the visual procedures, no algorithms were implemented to identify or accommodate curvilinear trend.

The software for data generation and evaluation was developed by the first author on the Macintosh Platform using RealBasic 4.02 and cross-compiled for Windows95/98NT. The accuracy of the program in implementing Nugent's (2000) visual analysis procedure was evaluated by checking program outputs against hand-calculations. The Monte Carlo Analyses were run on a Dell OptiPlex GX1 running Windows98.

In order to evaluate the visual procedures' ability to discern that no effect is present when there is indeed no effect present, random data streams were evaluated. For each level of N (total of 6 different N-sizes) and level of auto-correlation (total of 6 different levels), 1000 data streams were generated and the performance of the visual procedures and associated decision rules was evaluated (a total of 36,000 data streams).

Statistical convention dictates that a "good" procedure should indicate evidence of change when none is there no more than 5 (+/−2.5) percent of the time (type I error rate = .05). However, as Nugent (2000) points out,

> The use of the above rules and guidelines . . . will give the practitioner greater power to detect change. While this increased power may be bought at the price of an increase in type I error rates, this expense is more acceptable in a practice evaluation context than in a research context.

With respect to these considerations, the Monte Carlo analyses and results should be thought of as exploratory and descriptive. As such, we adopted no pre-set level of acceptable type I error rate in this study.

Table 1 shows the empirical type I error performance of Nugent's visual procedures for the 36,000 random data streams with no pro-

TABLE 1. Empirical Type I Error Rates of Nugent's Visual Procedure (Overall Type I Error and Individual Type I Errors for Each of the 4 Decision Rules) Under Various Levels of Autocorrelation (ρ) and Data Stream Length (N) with Equal Phase Lengths. 1000 Simulated Data Streams Were Evaluated per Condition (N and ρ)

N	ρ	Rule #1	Rule #2	Rule #3	Rule #4	Overall
20	−.4	.23	.08	.07	.40	.50
	−.2	.23	.07	.06	.40	.50
	0	.26	.10	.13	.44	.53
	.2	.29	.12	.15	.53	.61
	.4	.34	.18	.23	.58	.65
	.8	.51	.35	.41	.69	.78
16	−.4	.28	.09	.06	.29	.47
	−.2	.13	.10	.07	.31	.46
	0	.34	.13	.11	.37	.54
	.2	.38	.18	.18	.44	.59
	.4	.45	.23	.23	.49	.64
	.8	.60	.41	.40	.57	.74
14	−.4	.33	.11	.06	.25	.47
	−.2	.36	.13	.08	.30	.50
	0	.41	.17	.12	.34	.55
	.2	.44	.21	.16	.37	.59
	.4	.51	.28	.24	.41	.64
	.8	.63	.43	.37	.46	.73
12	−.4	.46	.17	.05	.29	.60
	−.2	.44	.17	.09	.33	.57
	0	.48	.22	.15	.35	.61
	.2	.53	.26	.18	.42	.67
	.4	.58	.33	.22	.46	.71
	.8	.68	.47	.34	.51	.79
10	−.4	.55	.24	.09	.38	.70
	−.2	.51	.22	.25	.39	.67
	0	.58	.26	.13	.42	.71
	.2	.65	.34	.37	.47	.77
	.4	.66	.35	.19	.50	.77
	.8	.74	.45	.24	.52	.81
8	−.4	.61	.24	.22	.42	.73
	−.2	.64	.30	.21	.46	.76
	0	.67	.31	.22	.48	.79
	.2	.71	.36	.24	.51	.81
	.4	.73	.42	.29	.58	.84
	.8	.76	.44	.34	.59	.85

grammed effects, using various N-sizes and levels of auto-correlation (ρ). The overall type I error rate is displayed as well as the type I error rates of each of the individual decision rules.

When faced with a limited number of data points and when auto-correlation is present in the data, a clinician using the visual procedures has up to an 85% chance of concluding that a change has occurred in patient-functioning when in fact one has not. The average type I error rate across various N-sizes and levels of auto-correlation is 66%. Rules #2 and #3 perform best by themselves (using the conventional type I error standard of .05) but only when no positive auto-correlation is present. Overall, the visual procedures proposed by Nugent (2000) do not adequately protect researchers and clinicians from inferring an effect when in fact there is none.

Next, the power (1 minus type II error) of the visual approach was evaluated. Given the high type I error rate, power for large effects should not be problematic. A more interesting investigation might be geared to evaluating how large a real effect has to be for the visual procedures to consistently detect it. An N-size of 14 (7 baseline and 7 treatment points) was selected (Crosbie, 1993, 1994 and Borckardt et al., under review) as it represents a balance between a reasonable amount of clinical data for evaluation and a realistic amount of data that a practicing clinician *could* collect (one data point per day for one week as a baseline following intake but prior to the first formal therapy session). However, there is some evidence to suggest that clinicians and clinical researchers are actually afforded fewer data points in real-world investigations (Huitema & McKean, 1991). Nonetheless, we have chosen 14 data points in order to be fair to Nugent's (2000) visual procedures with respect to power performance. Additionally, seven points per phase is the recommended minimum for other statistical procedures to provide adequate power and type I error control (Crosbie 1993, 1994; Robey et al., 1999). Various programmed effect-magnitudes were evaluated (.75, 1.5, 2.25, 3, 3.75, 4.5) using the visual procedures under different levels of auto-correlation (−.4, −.2, 0, .2, .4, .8).

The errors in the simulations followed the first order auto-regressive model:

$$e_t = \rho(e_{t-1}) + u_t$$

where t = 2 to N, ρ = programmed auto-correlation, e_1 = a normal random deviate N(0,1) generated via the polar method, and u_t = independ-

ent normal random deviate N(0,1) at time t also generated via the polar method. The data were generated as follows:

$$y_t = \alpha + e_t$$

where t = 1 to N, $\alpha = 0$ during baseline and various levels during the treatment phase (.75, 1.5, 2.25, 3, 3.75, 4.5; Crosbie, 1993, 1994). There were no programmed slopes in the data streams.

Statistical convention dictates that a "good" procedure for determining the presence of change should detect an effect when one is there at least 80% of the time (power = .80 = 1 minus type II error).

Table 2 shows the empirical performance of Nugent's visual procedures for 36,000 random data streams with various programmed between-phase effects (α), with a data stream length of 14 (seven points per phase) and various levels of auto-correlation (ρ). The overall power is displayed as well as the power of each of the individual decision rules.

Despite the very high type I error rates for the visual procedures, power is unacceptable with effect sizes of .75 and 1.5. While the overall visual procedure demonstrates acceptable power with an N of 14 and with effects as small as 2.25, the two decision rules that were identified as being the best performers with regard to type I error (rules #2 and #3) do not demonstrate adequate power with effects less than 4.5. Fortunately, the effect sizes in clinical time series studies tend to be large. Matyas and Greenwood (1990), for example, found the median effect size in their sample of clinical studies to be 10, with an effect size of 5 representing the 25th percentile.

To be appropriately discerning, a test must have acceptable type I error control and acceptable power. Unfortunately, Nugent's procedures do not satisfy the first requirement. However, a possible explanation is that the poor type I error control of the visual procedures might be related to the simplicity of the trend representation methods. As such, we examined a more conventional method for representing trend, namely, Ordinary Least Squares regression methods. The same methodology was implemented as above in terms of representing variability (sigma units) and evaluating treatment phase performance (use of decision rules). The only difference is that OLS methods were used to determine baseline-phase slope and intercept (and by implication, the extended trend line).

Table 3 shows the empirical performance of Nugent's visual procedures using OLS trend estimation methods for 36,000 random data

TABLE 2. Empirical Power of Nugent's Visual Procedure (Overall Power and Individual Powers for Each of the 4 Decision Rules) Under Various Levels of Autocorrelation (ρ) and a Data Stream Length of 14 with Equal Phase Lengths. Programmed Effect Sizes (α) Ranged from .75 to 4.5 in Increments of .75. 1000 Simulated Data Streams Were Evaluated per Condition (α and ρ). Acceptable Power Levels Are in Bold.

α	ρ	Rule #1	Rule #2	Rule #3	Rule #4	Overall
.75	−.4	.42	.20	.15	.40	.58
	−.2	.41	.20	.13	.38	.58
	0	.40	.21	.16	.40	.55
	.2	.46	.25	.21	.42	.62
	.4	.52	.31	.28	.45	.64
	.8	.67	.47	.41	.50	.77
1.5	−.4	.51	.29	.21	.51	.68
	−.2	.53	.30	.26	.53	.70
	0	.51	.31	.27	.54	.69
	.2	.57	.35	.34	.58	.74
	.4	.58	.39	.37	.55	.72
	.8	.70	.53	.48	.57	**.81**
2.25	−.4	.62	.43	.39	.70	**.80**
	−.2	.66	.46	.46	.71	**.82**
	0	.66	.48	.45	.71	**.82**
	.2	.68	.53	.50	.73	**.82**
	.4	.72	.54	.52	.70	**.84**
	.8	.74	.59	.54	.64	**.84**
3	−.4	.76	.62	.59	**.84**	**.89**
	−.2	.77	.62	.62	**.85**	**.91**
	0	.78	.65	.64	**.85**	**.90**
	.2	.77	.64	.65	**.85**	**.91**
	.4	.79	.68	.67	**.81**	**.90**
	.8	**.81**	.70	.65	.76	**.89**
3.75	−.4	**.83**	.72	.74	**.92**	**.95**
	−.2	**.86**	.73	.75	**.93**	**.97**
	0	**.88**	.79	**.81**	**.94**	**.97**
	.2	**.89**	.79	**.81**	**.91**	**.96**
	.4	**.87**	.79	.78	**.90**	**.96**
	.8	**.89**	**.80**	.75	**.82**	**.95**
4.5	−.4	**.91**	**.84**	**.85**	**.97**	**.99**
	−.2	**.91**	**.85**	**.88**	**.97**	**.98**
	0	**.92**	**.86**	**.87**	**.97**	**.99**
	.2	**.93**	**.88**	**.88**	**.95**	**.97**
	.4	**.87**	**.79**	**.78**	**.90**	**.96**
	.8	**.93**	**.89**	**.84**	**.87**	**.97**

TABLE 3. Empirical Type I Error Rates of Nugent's Visual Procedure Using OLS Trend Estimation Methods (Overall Type I Error and Individual Type I Errors for Each of the 4 Decision Rules) Under Various Levels of Autocorrelation (ρ) and Data Stream Length (N) with Equal Phase Lengths. 1000 Simulated Data Streams Were Evaluated per Condition (N and ρ).

N	ρ	Rule #1	Rule #2	Rule #3	Rule #4	Overall
20	−.4	.22	.04	.02	.21	.37
	−.2	.23	.06	.04	.29	.41
	0	.28	.08	.08	.33	.47
	.2	.39	.16	.16	.46	.60
	.4	.44	.24	.26	.51	.65
	.8	.68	.50	.53	.69	.82
16	−.4	.30	.09	.04	.18	.40
	−.2	.33	.09	.06	.26	.44
	0	.39	.14	.11	.33	.53
	.2	.43	.49	.16	.37	.56
	.4	.52	.27	.27	.47	.67
	.8	.67	.49	.47	.60	.78
14	−.4	.35	.10	.05	.22	.46
	−.2	.39	.12	.06	.25	.50
	0	.44	.18	.11	.29	.53
	.2	.50	.23	.18	.34	.58
	.4	.56	.32	.26	.42	.67
	.8	.72	.50	.43	.51	.78
12	−.4	.41	.15	.06	.27	.53
	−.2	.43	.15	.08	.32	.56
	0	.50	.22	.14	.38	.62
	.2	.55	.26	.19	.41	.65
	.4	.62	.35	.25	.45	.71
	.8	.75	.51	.41	.54	.82
10	−.4	.51	.21	.23	.37	.65
	−.2	.55	.21	.27	.40	.68
	0	.60	.28	.33	.47	.72
	.2	.64	.33	.38	.48	.76
	.4	.72	.40	.45	.52	.80
	.8	.80	.52	.55	.62	.86
8	−.4	.62	.27	.18	.45	.75
	−.2	.64	.33	.23	.53	.78
	0	.68	.34	.26	.52	.78
	.2	.74	.41	.33	.58	.82
	.4	.76	.45	.37	.64	.86
	.8	.86	.52	.41	.65	.89

streams with no programmed effects (i.e., for examination of type I error rates), with various N-sizes and levels of auto-correlation (ρ). The overall type I error rate is displayed as well as the type I error rates of each of the individual decision rules.

The results of this analysis do not indicate a distinct advantage of the use of OLS methods over Nugent's recommended trend estimation methods. It does not appear that the type I error problems associated with the visual procedure are due to the user-friendly, paper-and-pencil approach to trend estimation suggested by Nugent.

DISCUSSION

Nugent (2000) is very clear that the context of use for the visual procedures (practice evaluation) is not to be confused with that of statistical significance testing. However, simply because one is interested in practice evaluation does not justify lax standards. Inferring a treatment works when in fact it does not, is a disservice to patients and professionals alike. The medical and psychological literatures are replete with the tragic consequences of false hope [e.g., facilitated communication] (Mostert, 1995), the Mozart effect (McKelvie & Low, 2002) and therapeutic touch (Rosa, Rosa, Sarner & Barrett, 1998). The present study demonstrates that Nugent's visual approach is not an acceptable test for effect by any acceptable standard, research or otherwise. Far too frequently (two-thirds of the time) its application would lead clinicians and program evaluators to infer that the intervention made a difference, when, in fact, it did not.

Even in the context of clinical significance (as opposed to statistical significance) practitioners need to ask whether the patient has reached some pre-set criterion of mental health by chance alone. For example, if a patient's Global Severity Index (GSI) on the Symptom Checklist 90-R (a t-score with a mean of 50 and a StdDev of 10) changes from a baseline score of 68 (clinical range) to a score of 55 (normal range) after treatment onset, we would want to be careful about concluding that a change has occurred. In fact, we would be wrong to conclude that he/she has changed if the patient's baseline GSI variability is wrought with large swings of 15 points in either direction. In this case the correct decision about the presence of a real change would likely be that none has occurred. Nugent's (2000) visual procedures were specifically designed to handle just this type of scenario as the procedures involve rep-

resenting baseline trend and baseline variability. The treatment-phase data points are then evaluated against this important information. Unfortunately, the results of this study show that the visual procedures, despite the thoughtful design, too frequently result in incorrect decisions that effects are present when, in fact, they are not.

With regard to the individual decision rules, rule #1 (a significant change is indicated by a data point that falls more than 2 sigma units away from the extended trend line) appears to be the most likely to produce type I errors. Rule #4 (seven data points in a row that are on the same side of the extended trend line are indicative of significant change) is also too sensitive. Rules #2 and #3 (two of three data points falling more than 2 sigma units away from the extended trend line will signify significant change; and four of five data points on the same side of the extended trend line, and more than 1 sigma unit away from the extended trend line, signify significant change; respectively) appear to be the most discerning of the four. Unfortunately, rules #2 and #3 do not demonstrate good type I error control when positive auto-correlation is present, and they do not offer adequate power with effect sizes less than 4.5 standard deviations.

Effect sizes in clinical-case time series studies tend to be large, which might suggest that Nugent's visual procedures might be adequate if rules #1 and #4 are ignored. However, even conventional statistical methods (e.g., t-tests, ANOVA, regression) which have been deemed unusable for evaluating short clinical data streams due to unacceptable type I error control in the face of auto-correlation, have much better type I and type II properties than the visual procedures investigated in this study (Robey, 1999; Crosbie, 1993, 1994; Franklin, Allison & Gorman, 1994).

It does not appear that the type I error problems associated with the visual procedure are due to the simplistic, user-friendly approach to trend estimation. This is good news in that it may still be possible to use simple paper and pencil case-evaluation methods, if the procedures are modified. Specifically, one of the problems with the approach involves attempting to estimate baseline trend with very few available data points. This resulted in frequent modeling of trends that didn't actually exist during the Monte Carlo analyses and thereby resulted in poor decisions about the presence or absence of effects, especially using rule 4. When dealing with limited numbers of data points, decisions may often be more accurate regarding the presence or absence of change using approaches that assume a flat baseline (Borckardt et al., under review; Scruggs, Mastropieri & Casto, 1987). As such, it may be advantageous

to assume no baseline trend and to then test this assumption against patient report. If indeed the patient reports that symptoms appear to be consistent (no trend), a simple visual analytic approach that assumes flat baselines could be applied (see Scruggs, Mastropieri & Casto, 1987). If a patient verifies the presence of a baseline trend, Nugent's (2000) procedure may very well work appropriately without modification (however, this has yet to be tested).

While Nugent's (2000) visual procedure is a noble effort, it is presently of little use in helping clinicians decide whether their treatment worked or not. The procedure is too likely to report that a change has occurred when, in fact, it has not. Additionally, if small but real effects are present, they may be missed by the application of the procedures. As such, investigators who are interested in discerning the presence or absence of real change in clinical-case time series data streams, are encouraged to look elsewhere for tools to aid them (see Scruggs, Mastropieri & Casto, 1987; Borckardt et al., under review; Ferron, 1990; Franklin, Allison & Gorman, 1990).

REFERENCES

Beutler, L. E. (2000). David and Goliath: When empirical and clinical standards of practice meet. *American Psychologist, 55* (9), 997-1007.

Bloom, M., Fischer, J., and Orme, J. G. (2003). *Evaluating practice: Guidelines for the accountable professional*, fourth edition. Boston: Allyn & Bacon.

Borckardt, J. J., Murphy, M. D., Nash, M. R., Moore, M., and Shaw, D. (under review). *A re-evaluation and modification of the PND-statistic for use with very short data streams.*

Borckardt, J. J., and Nash, M. R. (2002). How practitioners (and others) can make scientifically viable contributions to clinical-outcome research using the single-case time series design. *International Journal of Clinical & Experimental Hypnosis, 50* (2), 114-148.

Callaghan, G. M. (2001). Demonstrating clinical effectiveness for individual practitioners and clinics. *Professional Psychology: Research and Practice, 32* (3), 289-297.

Clement, P. W. (1994). Quantitative evaluation of 26 years of private practice. *Professional Psychology: Research and Practice, 25* (2), 173-176.

Crosbie, J. (1993). Interrupted time-series analysis with brief single subject data. *Journal of Consulting and Clinical Psychology, 61* (6), 966-974.

Crosbie, J. (1994). Interrupted time-series analysis with short series: Why it is problematic; how it can be improved. In J. M. Gottman (Ed). *The analysis of change.* NJ: Erlbaum.

Ferron, J., and Onghena, P. (1996). The power of randomization tests for single-case phase designs. *The Journal of Experimental Education, 64* (3), 231-239.

Ferron, J., and Sentovich, C. (2002). Statistical power of randomization tests used with multiple-baseline designs. *The Journal of Experimental Education, 70* (2), 165, 178.

Franklin, R. D., Allison, D. B., and Gorman, B. S. (1996). *Design and analysis of single-case research*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Gedo, P. M. (1999). Single case studies in psychotherapy research. *Psychoanalytic Psychology, 16* (2), 274-280.

Goldfried, M. R., and Wolfe, B. E. (1998). Toward a more clinically valid approach to therapy research. *Journal of Consulting and Clinical Psychology, 66* (1), 143-150.

Gottman, J. M. (1981). *Time-series analysis: A comprehensive introduction for social scientists*. Cambridge, UK: Cambridge University Press.

Hilliard, R. B. (1993). Single-case methodology in psychotherapy process and outcome research. *Journal of Consulting and Clinical Psychology, 61* (3), 373-380.

Huitema, B. E., and McKean, J. W. (2000). A simple and powerful test for autocorrelated errors in OLS intervention models. *Psychological Reports, 87*, 3-20.

Huitema, B. E., and McKean, J. W. (1991). Auto-correlation estimation and inference with small samples. *Psychological Bulletin, 110* (2), 291-304.

Knuth, D. E. (1981). *The art of computer programming: Vol. 2. Semi-numerical algorithms* (2nd ed.). Reading, MA: Addison-Wesley.

Lowman, R. L. (2001). Constructing a literature from case studies: Promise and limitations of the method. *Consulting Psychology Journal: Practice and Research, 53* (2), 119-123.

Marten, P. A., and Heimberg, R. G. (1995). Toward an integration of independent practice and clinical research. *Professional Psychology: Research and Practice, 26* (1), 48-53.

Matyas, T. A., and Greenwood, K. M. (1990). Visual analysis of single-case time-series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*, 341-351.

McKelvie, P., and Low, J. (2002). Listening to Mozart does not improve children's spatial ability: Final curtains for the Mozart effect. *British Journal of Developmental Psychology, 20* (2), 241-258.

McKnight, S. D., McKean, J. W., and Huitema, B. E. (2000). A double bootstrap method to analyze linear models with auto-regressive error terms. *Psychological Methods, 5* (1), 87-101.

Mostert, M. P. (2001). Facilitated communication since 1995: A review of published studies. *Journal of Autism & Developmental Disorders, 31* (3), 287-313.

Nugent, W. R. (2000). Single case design visual analysis procedures for use in practice evaluation. *Journal of Social Service Research, 27* (2), 39-75.

Ostrom, C. W. (1991). Time series analysis: Regression techniques, second edition. Series: *Quantitative applications in the social sciences*. Series Editor: Michael S. Lewis-Beck. Sage Publications, Inc.

Rosa, L., Rosa, E., Sarner, L., and Barrett, S. (1998). A close look at therapeutic touch. *Journal of the American Medical Association, 279* (13), 1005-1010.

Schwartz, R. M. (1997). Consider the simple screw: Cognitive science, quality improvement, and psychotherapy. *Journal of Consulting and Clinical Psychology, 65* (6), 970-983.

Scruggs, T. E., Mastropieri, M. A., and Casto, G. (1987). The quantitative synthesis of single subject research: Methodology and validation. *Remedial and Special Education, 8* (2), 24-33.

RECEIVED: 01/03
REVISED: 04/03
ACCEPTED: 05/03