COMMENTS ON "A SIMPLIFIED TIME-SERIES ANALYSIS FOR EVALUATING TREATMENT INTERVENTIONS"

CAROL JOYCE BLUMBERG

UNIVERSITY OF DELAWARE

In a recent article, Tryon (1982) introduced several hypothesis testing procedures, based on the use of the *C* statistic (Young, 1941), for the analysis of data from several phase time-series designs such as the baseline phase-treatment phase design. These procedures can be summarized as follows:

Procedure 1. The *C* statistic can be used to test whether or not there is a trend within one particular phase, usually the baseline (Young, 1941).

Procedure 2. When the baseline phase is found under Procedure 1 to show no statistically significant trend, the *C* statistic can then be used to test whether there is a trend in the data when the baseline and treatment phases are treated as an aggregate set of time-series data. If a trend is found using this procedure, it is concluded that the treatment has been effective (Tryon, 1982).

Procedure 3. When the baseline phase is found under Procedure 1 to show a statistically significant trend, the *C* statistic can then be used to test whether this trend continues in the treatment phase. Here, the null hypothesis is that the trend continues and the alternative hypothesis is that the trend changes. If a change in trend is found using this procedure, it is concluded that the treatment has been effective (Tryon, 1982). When implementing Procedure 3, the data analyst is required to assume that the trend in the baseline data is of some particular shape (e.g., linear, quadratic, or logarithmic).

The *C* statistic is computed slightly differently under each of the procedures, and then is transformed into a *z* statistic. *Note:* The Tryon (1982)

article contains two typographical errors. In Equation 3, the denominator of the fraction should be

$$2 \cdot \sum_{i=1}^N (X_i - \bar{X})^2.$$

Equation 4 should read

$$S_c = \sqrt{\frac{N - 2}{(N - 1) \cdot (N + 1)}}.$$

Young (1941) has shown that under Procedure 1, *z* has approximately a standard normal distribution when $N \geq 20$. When $N \leq 25$, Young has calculated the critical values for *z* for both $\alpha = .05$ and $\alpha = .01$. When $\alpha = .05$, the critical value is 1.65 for all $N \leq 25$. For $\alpha = .01$, Young's table should be consulted (Tryon, 1982; Young, 1941). It should be noted that when using Young's table implicitly, a one-tailed test is being performed, because Young assumes that only high-enough positive values of *C* (and hence *z*) should be considered statistically significant. Young's distributional results can easily be extended to cover Procedures 2 and 3.

The purpose of this article is to describe some of the problems that can be encountered when these procedures are used.

PROBLEMS

Tryon (1982) identified two problems associated with Procedure 3:

1. When a linear baseline phase trend is compared to a linear treatment phase trend having a different slope, the *C* statistic will have the same value regardless of the actual values of the slopes of the baseline and treatment phase lines. This value of *C* is totally dependent upon the number of observations taken. Hence, the number of ob-

I am grateful to George Brabner for his helpful comments.

Reprint requests should be sent to Carol Joyce Blumberg, Department of Educational Studies, University of Delaware, Newark, Delaware 19716.

Figure 1. Graph of $X_i = (i)^2$, (a); a slight treatment effect (b); and a sizable treatment effect (c).

lem is. When $\alpha = .05$, a minimum of only four baseline and four treatment phase observations are needed to achieve statistical significance. When $\alpha = .01$, a minimum of only six baseline and six treatment phase observations are needed to achieve statistical significance.

2. When there are linear trends in both the baseline and treatment phases with the same slopes but with different intercepts (i.e., the treatment phase observations are shifted either upward or downward by a constant amount from the baseline phase observations), the C statistic will not be significant, even though the treatment has had an effect. But, Tryon's statement is much too narrow. In fact, if the baseline and treatment phases have trends of the same shape with the treatment phase shifted either upward or downward by some constant amount from the baseline phase, the C statistic will not pick up this shift between the baseline and treatment phases, no matter what this common shape is. For example, the C statistic cannot distinguish between the data sets presented in Figure 1a, b, and c. Yet, in panel a there is no treatment effect; in panel b there is a slight effect; and in panel c there is a sizable effect.

There are additional problems with the use of the C statistic:

3. It is often possible to get a statistically significant z value under Procedure 3 even when there is no change in the trend of the data between the baseline phase and the treatment phase. For example, if the values of the X_i s during both the baseline and treatment phases are determined by the equation $X_i = (i)^2$, as in Figure 1a, and four observations are taken during both the baseline and the treatment phases (i.e., eight observations in all), then Procedure 1 performed on the baseline observations yields a z value of 1.86, which is statistically significant at the $\alpha = .05$ level. Hence, the data analyst would correctly conclude that there is a trend in the baseline data. Procedure 3 is then used to test whether the trend continues into the treatment phase and, in this example, yields a z value of 1.92 ($p < .05$). Hence, the data analyst will wrongly conclude that there was a change in the trend. This example is not a rarity; rather it is indicative of what will often happen for many dif-

servations taken, rather than the trends themselves (as should be the case), will determine whether a statistically significant trend is found. Tryon, however, does not point out just how serious this prob-

ferent types of trends in the data whenever enough observations are taken during both the baseline and the treatment phases and no treatment effects are present.

4. Under Procedure 1, because of power considerations, it is often possible to conclude when a small number of observations are taken that the baseline data show no trend when, in fact, there is a trend. If it is concluded from Procedure 1 that there is no trend in the baseline data, then Procedure 2 is used to determine whether there is a trend in the treatment phase data, even if the conclusion from Procedure 1 is incorrect. Under Procedure 2, even when the trend remains unchanged from the baseline phase through the treatment phase, this trend will now invariably be shown to be statistically significant solely because of the increased number of observations, which have increased the power of the test. To illustrate this phenomenon, again let $X_i = (i)^2$ and take three observations during the baseline phase and only two additional observations during the treatment phase. Starting with the baseline observations, Procedure 1 yields a z value of 1.36 ($p > .05$). Because this z value is not statistically significant, it is wrongly concluded that there is no trend in the baseline data, and Procedure 2 is inappropriately applied to test whether there is a trend in the treatment phase. Procedure 2 yields a z value of 2.21 ($p < .01$), and it is concluded that there is a trend (and hence a treatment effect) in the treatment phase data. Consequently, a second wrong conclusion has been made.

This process of making two wrong conclusions, when there is a trend in the baseline data that continues into the treatment phase, will happen almost invariably as long as a fairly small number of observations are taken during the baseline phase. Because one cannot estimate a priori how many baseline observations are needed before Procedure 1 will detect the trend in the baseline phase, a serious flaw exists in the procedures recommended by Tryon. Although most of the variables used in single-subject research have stable baselines, there are many examples of variables for which there is a trend in the baseline. Some are: (a) caloric intake for persons with anorexia nervosa (Agras, Barlow,

Chapin, Abel, & Leitenberg, 1974) has a decreasing linear trend; (b) time and distance walked by an agoraphobic patient (Agras, Leitenberg, & Barlow, 1968) has an increasing linear trend; (c) school attendance (Barber & Kagey, 1977) has a quadratic trend; and (d) points earned by patients with depression (Eisler & Hersen, 1973; Hersen, Eisler, Alford, & Agras, 1973) has a logarithmic trend.

Problems 3 and 4 are symptomatic of the major shortcoming of the C statistic under Procedures 2 and 3. The methods suggested under Procedures 2 and 3 are really conditional testing situations. That is, the data analyst must always first perform the test described under Procedure 1. The results of Procedure 1 are then inspected; based on these results, it is decided whether to perform the second test using Procedure 2 or Procedure 3. Procedure 2 is used if Procedure 1 yields a statistically non-significant z , whereas Procedure 3 is used if Procedure 1 yields a statistically significant z value. Hence, which procedure is used as the second test is dependent (i.e., conditional) upon the results from the first test performed using Procedure 1. It is well known that in conditional testing situations such as this, Type I (α) and Type II (β) errors are inflated, sometimes greatly. Hence, data analysts may think that they are using $\alpha = .05$ under Procedure 2 or 3, but α may be, for example, as high as .80.

CONCLUSION

Several problems have been discussed with respect to the procedures suggested by Tryon (1982). Because of the severity of the problems, especially the conditional testing problem and problems 3 and 4, it is my opinion that Procedures 2 and 3 should never be used. Still, Procedure 3 is an important step in the right direction because it requires the data analyst to take into account the shape of the baseline trend. One method for analyzing single-subject data that takes into consideration the shape of the baseline data is a least squares approach (either linear or non-linear regression depending on the shape of the baseline data). This method, however, often leads to large standard errors.

Several procedures have emerged from recent literature on the analysis of longitudinal quasi-experimental data that correctly test for treatment effects in several phase designs (Blumberg, 1983; Bryk, Strenio, & Weisberg, 1980; Bryk & Weisberg, 1976; Olejnik, 1978; Strenio, Weisberg, & Bryk, 1983). All of these procedures explicitly take into account the shape of the baseline trend and often lead to reasonably sized standard errors. Further, these authors have shown that the shape assumed (or lack of shape assumed) for the trend in the baseline data can have a major influence on the estimates obtained for treatment effects. If the wrong shape is assumed, grossly inaccurate estimates will be obtained. But, all of the above-cited procedures require observations on two or more subjects at each time point where data are collected. Perhaps, however, these procedures can be modified to accommodate single-subject designs.

REFERENCES

- Agras, W. S., Barlow, D. H., Chapin, H. N., Abel, G. G., & Leitenberg, H. (1974). Behavior modification of anorexia nervosa. *Archives of General Psychiatry*, **30**, 279-286.
- Agras, W. S., Leitenberg, H., & Barlow, D. H. (1968). Social reinforcement in the modification of agoraphobia. *Archives of General Psychiatry*, **19**, 423-427.
- Barber, R. M., & Kagey, J. R. (1977). Modification of school attendance for an elementary population. *Journal of Applied Behavior Analysis*, **10**, 41-48.
- Blumberg, C. J. (1983). The estimation and hypothesis testing of treatment effects in nonequivalent control group designs when continuous growth models are assumed (Doctoral dissertation, Michigan State University, 1982). *Dissertation Abstracts International*, **43**, 2952B.
- Bryk, A. S., Strenio, J. F., & Weisberg, H. I. (1980). A method for estimating treatment effects when individuals are growing. *Journal of Educational Statistics*, **5**, 5-34.
- Bryk, A. S., & Weisberg, H. I. (1976). Value-added analysis: A dynamic approach to the estimation of treatment effects. *Journal of Educational Statistics*, **1**, 127-155.
- Eisler, R. M., & Hersen, M. (1973, August). *The A-B design: Effects of token economy on behavioral and subjective measures in neurotic depression*. Paper presented at the meeting of the American Psychological Association, Montreal, Quebec.
- Hersen, M., Eisler, R. M., Alford, G. S., & Agras, W. S. (1973). Effects of token economy on neurotic depression: An experimental analysis. *Behavior Therapy*, **4**, 392-397.
- Olejnik, S. F. (1978). Data analysis strategies for quasi-experimental studies where differential group and individual growth rates are assumed (Doctoral dissertation, Michigan State University, 1977). *Dissertation Abstracts International*, **39**, 200A.
- Strenio, J. F., Weisberg, H. I., & Bryk, A. S. (1983). Empirical Bayes estimation of individual growth curve parameters and their relationship to covariates. *Biometrika*, **39**, 71-86.
- Tryon, W. W. (1982). A simplified time-series analysis for evaluating treatment interventions. *Journal of Applied Behavior Analysis*, **15**, 423-429.
- Young, L. C. (1941). On randomness in ordered sequences. *Annals of Mathematical Statistics*, **12**, 293-300.

Received September 29, 1983

Final acceptance May 3, 1984