JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1974, 7, 629-633 NUMBER 4 (WINTER 1974)

A FURTHER CONSIDERATION IN THE APPLICATION OF AN ANALYSIS-OF-VARIANCE MODEL FOR THE INTRASUBJECT REPLICATION DESIGN

T. KRATOCHWILL, K. ALDEN, D. DEMUTH, D. DAWSON, C. PANICUCCI, P. ARNTSON, N. MCMURRAY, J. HEMPSTEAD, AND J. LEVIN[1]

UNIVERSITY OF WISCONSIN, MADISON

It is argued that the analysis-of-variance model is inappropriate for assessing treatment effects in single-subject designs. In particular, such designs are demonstrated to violate the crucial assumption concerning the statistical independence of observations. Alternative methods of data analysis are suggested.

Over the years, there has been considerable discussion about the use of single-subject (N = 1) designs to observe behavioral changes within the operant conditioning paradigm. One area of discussion has been concerned with the desire of some researchers to justify their results by applying inferential statistics to N = 1 designs. This desire has resulted in a recent endeavor to adapt analysis-of-variance (ANOVA) procedures to N = 1 data (Gentile, Roden, and Klein, 1972) under the assumption that the subject may be regarded as "... a response generator the responses of which to a particular stimulus are statistically independent and normally distributed about a central response value" (Shine and Bower, 1971, p. 112). Our purpose is to demonstrate that the statistical independence assumption is entirely unwarranted in an N = 1 design and as a result, the recently suggested ANOVA procedure is not appropriate for such data (*cf.* Glass, Peckham, and Sanders, 1972).[2]

Because simple inspection of the data from Klein's (1971) study did not reveal "dramatic changes", Gentile *et al.* (1972) attempted to develop a theoretical rationale for the application of a one-way ANOVA to the data. In their analysis, treatment phases (A and B) represent the traditional "between" source of variance and the single subject's within-phase performance, consisting of N observations in which the subject was judged as being "on" (1) or "off" (0) task, represents the "within" source of variance. The subject's mean number of on-task responses within a given treatment phase is then used as an estimate of its "true" mean for that treatment. Such a procedure is reasonable. Our major objection focuses on the authors' claim that each of the subject's N responses entering into the treatment mean—which comprise the experimental units in the analysis that follows—are mutually independent, an assumption necessary for the analysis they advocate. We similarly object to Gentile *et al.*'s argument that a subject's sequence of "on-task" responses under a given treatment may be likened to a sequence of "head", "tail" outcomes in a coin-tossing experiment.

[1]With the exception of the first and last authors, the order of authorship was determined randomly. We are grateful to Roger Klein for making his procedures available to us and to Roger Severson and Michael Subkoviak for their valuable inputs to the paper.

This is one in a series of articles available for $1.50 from the Business Manager, *Journal of Applied Behavior Analysis,* Department of Human Development, University of Kansas, Lawrence, Kansas 66045. Ask for Monograph #4.

[2]In this paper, we have chosen not to address ourselves to other aspects of the Gentile *et al.* (1972) article with which we take issue. This includes the authors' claim that the use of an ABAB (or similarly permuted) design circumvents the problems associated with correlated data, a claim that is based on an erroneous interpretation of the assumptions required for within-subject ANOVA.

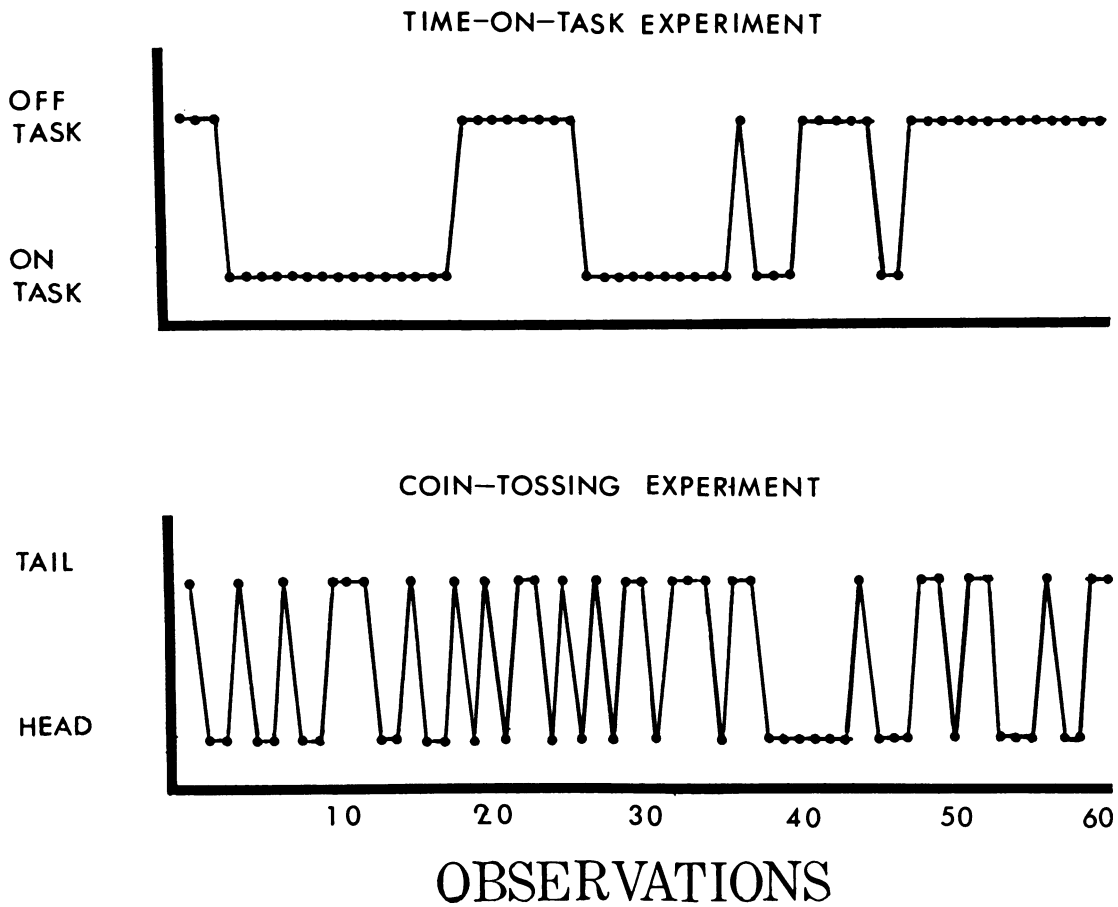

Fig. 1. A comparison of the pattern of consecutive observations in the two experiments.

Are the Observations in an $N = 1$ Experiment Statistically Independent?

Since most of us are of the persuasion that "actions speak louder than words", we conducted the necessary coin-tossing and time-on-task "experiments", making the prediction that the usual pattern of time-tied correlations would be obtained in the latter but not in the former experiment. That is to say, in the time-on-task experiment, we expected to find that temporally adjacent observations would be highly correlated, whereas in the coin-tossing experiment we expected that this correlation would approach zero (thereby fulfilling a necessary requirement of statistical independence).

In the time-on-task experiment, a 10-yr-old child was videotaped during a typical classroom routine. The videotape was then independently observed by two school psychologists over a 15-min period. Recording procedures were identical to those of Klein (1971) in which the subject was scored as being either "on" or "off" task for each observation.[3] By coincidence, an intact 10-min period containing 60 observations occurred during the 15-min observational period where the subject was on task exactly 50% of the time. This time-slice was selected in order to simplify the coin-tossing analogy (*i.e.*, if this child were happened upon during the course of an experiment, a reasonable estimate of his base-

[3]Interrater reliability was calculated by a per cent agreement method in which the number of agreements was divided by the total number of time intervals. The computed reliability was 98% for the two observers.

line on-task probability would be given by $p = \frac{1}{2}$, while the probability of obtaining a head on a given flip of a fair coin is also represented by $p = \frac{1}{2}$). Thus, we performed a second experiment in which a computer simulated 60 independent tosses of a fair coin. Out of 60 "tosses", 32 heads were obtained.[4]

A visual comparison of the two experiments is provided in Figure 1. Note that in the time-on-task experiment, there are several strings of consecutively identical responses, while in the coin-tossing experiment the sequence is more nonsystematic (*e.g.*, randomly up and down). A simple way to compare the two experiments is in terms of the number of times the sequence changes from one outcome to another over time (*i.e.*, from "off task" to "on task" or from "tail" to "head"). In the time-on-task experiment, there are eight such changes of direction, whereas in the coin-tossing experiment there are 34 such changes. Under the assumption that the two experiments yield analogous results, the chances that 42 changes of direction would be split 34 to eight (or more extreme) are less than one in a thousand.

There is another way in which the results of the two experiments may be compared. After adding one more observation to each experiment (yielding a total of $N = 61$ observations in each), 2×2 contingency tables were constructed that contained the number of times a particular outcome on observation t was preceded by a particular outcome on observation t-1. For example, the upper left-hand entries in the two experiments of Table 1 represent the number of times that an "on-task" response (or a "head") on one observation followed an "on-task" response (a "head") on the preceding observation. The information contained in Table 1 is, therefore, based on the 60 pairs of adjacent observations in each experiment.

Note that in the time-on-task experiment, the subject's behavior was recorded as being the same on two consecutive observations (either "on task", "on task", or "off task", "off task") in 52 of 60 cases, 87% of the time. The corresponding entries for the coin-tossing experiment result in a total of 25 of 60, 42% (with the expected result, assuming statistical independence, being 30 of 60, 50%). A correlation coefficient based on the time-on-task data is given by $\phi = 0.73$, while in the coin-tossing experiment, $\phi = -0.17$, actually in the opposite direction (as compared with the theoretical value of zero). The large correlation associated with the time-on-task experiment indicates that the subject's behavior was far more likely to be recorded as being the same on two consecutive observations than as being different.

Table 1

A comparison of the relationship between consecutive observations in the two experiments.

A. Time-On-Task Experiment

Observation t-1	Observation t: *On Task*	Observation t: *Off Task*
On Task	26	4
Off Task	4	26

B. Coin-Tossing Experiment

Observation t-1	Observation t: *Head*	Observation t: *Tail*
Head	15	17
Tail	18	10

Although other measures could be used to compare the results of the two experiments, the ones already discussed permit conclusions about statistical independence assumptions. That is, since the correlational patterns differ greatly in the two experiments, Gentile *et al.*'s (1972) basis for equating $N = 1$ and coin-flipping experiments is empirically unfounded. Under the assumptions of $p = \frac{1}{2}$, statistical independence,

[4]It should be mentioned that the general arguments that follow do not depend on the particular value of p selected.

and 60 pairs of adjacent observations, each cell in Table 1 would be expected to contain 15 entries; in the coin-tossing experiment, the entries are clearly more compatible with what was expected (they range from 10 to 18) than they are in the time-on-task experiment (range = four to 26).[5]

Although the probability of obtaining a head on a particular toss of a coin is unrelated to the outcomes immediately preceding it, the probability that a subject is on task on a given observation is not unrelated to his immediately prior behavior. Neither do we know of one general theory of human behavior that would support the assumption that what a person does at time t is totally independent of what he did at time t-1, t-2, t-3, *etc.* The relatedness of consecutive human behaviors is nicely illustrated in one of the important behavior change techniques employed by behavior modifiers, that of shaping, where true independence would preclude the success of the technique. If consecutive human behaviors were in fact independent, it would not be possible to develop the procedure of shaping, based on the gradual change of behaviors over consecutive trials.

A Footnote

Our rejoinder to Gentile *et al.* (1972) has focused exclusively on the assumption that responses generated in an N = 1 design could be treated as being statistically independent. If such an assumption were tenable, the use of univariate ANOVA procedures might be appropriate. It is beyond the scope of this note to offer detailed accounts of alternative statistical analyses that might be considered in (N = 1) designs where independence assumptions are not warranted. Certainly, statistical techniques in which the subject's correlational pattern is given its due consideration would seem appropriate. While it is hard to envision multivariate procedures applied to N = 1 designs, the recent developments in time-series analysis offer a great deal of promise in this domain (Glass, Wilson, and Gottman, 1972; Gottman, McFall, and Barnett, 1969).

Apart from the time-worn suggestion that behavior modifiers make use of simple graphical plots of their data to determine whether or not certain predicted outcomes occurred, we would add that researchers might consider the use of nonparametric and randomization models in conjunction with such plots. These could range from easy-to-compute sign tests reflecting predicted increases and decreases of on-task behavior in different phases of an ABAB design, to more sophisticated rank tests in which *a priori* trend or contrast coefficients are applied to the data.

If the goal of the researcher is to demonstrate that a particular predicted pattern of results was statistically confirmed, such procedures could be employed. It is our belief, however, that the researcher's goal *should be* to decide whether or not the subject that he is treating has reached a response criterion that is considered to be meaningful for that subject. In such cases, a simple tally of responses and/or plot of the data would suffice in making practical decisions about that subject. To talk about a "statistically" significant change within a subject makes no sense to us without regard both to its "practical" significance and to questions of internal, and especially external, validity (Bracht and Glass, 1968; Campbell and Stanley, 1966).

[5]While we do not wish to argue on the basis of these two N = 1 experiments that the *particular* results are generalizable to other subjects or coins, we are arguing that the comparative *patterns* of results are generalizable.

REFERENCES

Bracht, G. H. and Glass, G. V. The external validity of experiments. *American Educational Research Journal,* 1968, **5**, 437-474.

Campbell, D. T. and Stanley, J. C. *Experimental and quasi-experimental designs for research.* Chicago: Rand-McNally, 1966.

Gentile, J. R., Roden, A. H., and Klein, R. D. An analysis-of-variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis,* 1972, **5**, 193-198.

Glass, G. V., Peckham, P. D., and Sanders, J. R. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research,* 1972, **42,** 237-288.

Glass, G. V., Wilson, V. L., and Gottman, J. M. *Design and analysis of time-series experiments.* Boulder: Laboratory of Educational Research, University of Colorado, November 1972.

Gottman, J. M., McFall, R. M., and Barnett, J. T. Design and analysis of research using time series. *Psychological Bulletin,* 1969, **72,** 299-306.

Klein, R. D. *The effects of a systematic manipulation of contingencies upon overt work behavior in a primary classroom.* Unpublished doctoral dissertation, State University of New York at Buffalo, 1971.

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

Received 17 January 1973.
(Published without revision.)