JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1972, 5, 193-198 NUMBER 2 (SUMMER 1972)

AN ANALYSIS-OF-VARIANCE MODEL FOR THE INTRASUBJECT REPLICATION DESIGN[1]

J. RONALD GENTILE, AUBREY H. RODEN, AND ROGER D. KLEIN

STATE UNIVERSITY OF NEW YORK AT BUFFALO AND UNIVERSITY OF PITTSBURGH

One- and two-way analysis-of-variance procedures are shown logically to be appropriate for testing hypotheses in successive treatment reversal designs for one-subject and N-subject experiments, respectively. The applicability of these designs is demonstrated through analyses of typical data.

The preponderance of studies conducted within the paradigm of operant behavior employ the intrasubject replication design (often called, simply, the reversal design), in which various treatments are successively applied to and removed from the same subject (Sidman, 1960). For example, the A-B-A-B design (in which A = baseline of no reinforcement for a certain response, B = the contingent availability of a reinforcing stimulus following that response) is widely used to demonstrate that if a certain reinforcer is made available contingent upon a response (treatment B), the effect is to increase the frequency of that response above operant level (condition A). Then conditions are reversed and treatment A is reinstated, during which time the response rate is expected to revert to its operant level. Finally, treatment B is re-applied and the response rate is expected once again to increase to above operant level.

Control over the behavior in question can be said to be obtained only when the B treatment provides a *significant change* in response rate over that obtained in A. In other words, changes in behavior must reliably occur as a result of these treatment changes or else the investigator cannot infer a functional relationship between the treatment and the behavior. Also, such changes in behavior as a result of changes in the treatment conditions provide the most convincing demonstration of functional relationships.

The importance of this general procedure for the experimental analysis of behavior can hardly be overemphasized. This design, or variants of it, has been the vehicle for many principles of behavior developed in the last several decades and for most of the successes of operant behavior modification procedures in practical settings. (Indeed, the general form of this argument—if A, then B; if not A, then not B—is one of the most fundamental arguments in scientific methodology.)

Nevertheless, there are some disadvantages of the reversal design (see Bandura, 1969, pp. 242-244), one of the most serious of which is the interpretive problem of how large does the behavior change from treatment to treatment have to be to be considered a significant change. As Bandura points out, interpretation is not difficult provided that large successive behavior changes occur rapidly and consistently for many subjects. The interpretive problem arises in those cases in which the behavioral changes are not dramatic or in which some individuals remain unaffected by repeated exposure to one of the treatments. Such findings are especially prevalent in situations, such as classrooms, in which laboratory controls for creating favorable experimental conditions are difficult to achieve. The problem as stated reduces to a statistical one:

[1]Appreciation is expressed to S. David Farr, Malcolm J. Slakter, Thomas J. Shuell, Richard Spencer and Kevin Crehan for their assistance. Reprints may be obtained from J. Ronald Gentile, Educational Psychology Dept., State University of New York, Buffalo, N.Y. 14214.

> "... no statistical criteria have been developed to evaluate whether the magnitude of change produced by a given treatment exceeds the variability resulting from uncontrolled factors operating while the treatment condition is not in effect." (Bandura, 1969, p. 243)

Probably the major reason for the lack of concern with statistical analyses of the data is the assumed inapplicability of statistical techniques to individual cases or small numbers of subjects. Properly conceived, however, studies that collect repeated observations on the same individual (as is typical of most operant studies using time-sampling techniques or of clinical studies collecting data on patients) are appropriate for the analysis of variance model. The purpose of the present paper is to provide the rationale behind the use of an analysis-of-variance model for the reversal design, both for the one subject case and for more than one subject.

A PROBLEM

Klein (1971) investigated the effects of teacher attention and tokens (stars, which could later be exchanged for time spent on activities during a play period) contingent upon on-task behavior, compared with tokens contingent upon task completion, on three kindergarten students. On-task behavior was defined as the student attending to the task for a specified period of time, while task completion was correctly finishing the task. One expectation was that making reinforcement contingent upon the less-demanding on-task response would increase the time these students spent in on-task behavior. The phases of the study follow:

- A_1: baseline, in which a token economy was in operation, reinforcement of five tokens being contingent upon task completion.
- B: teacher attention was made contingent upon on-task behavior, the teacher approaching the child periodically when he was working on the task and commenting favorably on his behavior. Five tokens were still received at the time of task completion.
- C_1: tokens were made contingent upon on-task behavior, the teacher approaching the child periodically as in Phase B but, in addition to supplying social reinforcement, supplying tokens. There were still five tokens per task, but they were now distributed throughout the task.
- A_2: return to baseline.
- C_2: return to condition C_1.

It is beyond the scope of this paper to go into detail on the methodology of Klein's study, since it is being discussed only as a vehicle for understanding the need for the model we present below. Suffice it to say that conditions A, B, and C constitute three experimental treatments, the independent variable. The measured result of the manipulation, the dependent variable, is a binomial distribution obtained by observers using a time-sampling procedure, *i.e.,* number of on-task responses.

The results are indicated in Table 1, and they provide some difficulties in interpretation. For instance, although there are differences as predicted among treatments A_1, B, C_1, and A_2 for James, are these differences reliable? Why did the return to condition C (Phase C_2) not have the effect of increasing the number of on-task behaviors over baseline as it had in Phase C_1? Another problem is that the same profile of results is not shown by Lynn, since on-task behavior in C_2 was increased over baseline and, more to the point, since it was increased in B over C.

Differences such as these are not easily rationalized by inspection of the data. On the other hand, 24 days of actual classroom time were involved in the collection of these data, not including the planning time, training the teacher and motivating her for the extra work, *etc.* A

Table 1

Mean proportions, standard deviations, and number of observations (recorded intervals) of on-task behavior for each phase for two students from Klein (1971).

Student		*Phase* A_1	B	C_1	A_2	C_2
James	Mean Proportion	0.286	0.331	0.372	0.219	0.202
	Standard Deviation	0.452	0.471	0.484	0.415	0.402
	No. of Observations	974	610	529	210	282
Lynn	Mean Proportion	0.260	0.446	0.381	0.266	0.301
	Standard Deviation	0.439	0.498	0.486	0.445	0.460
	No. of Observations	1023	480	565	90	143

statistical technique would certainly be useful to aid the experimenter in making sense of the data and to suggest whether replication would be profitable.

THE ONE-WAY ANALYSIS-OF-VARIANCE MODEL

We propose that the experiment for each subject discussed above, and other experiments similar to it, be conceived as a one-way analysis-of-variance design, with (1) treatment effects being what is traditionally considered the between-subjects effects, and (2) number of observations being considered the standard within-subjects effects. With this change in conception, the justification of which follows, traditional formulas can be used to test the hypothesis that behaviors arising from each treatment condition could have been drawn by chance from the same population.

Behavior of a single person can be conceived as a chain of response events occurring in time. Any given response of interest, adequately defined (such as the number of sneezes or vocalizations, per cent of time spent reading, *etc.*), can be viewed as occurring with some frequency per specified time period. The average number of responses over some large number of such time periods can be considered to approach the "true" frequency distribution of that response for that person. Further, each response can be considered as independent of every other response in the same class. Then, an unbiased estimate of the response belonging to the defined class, obtained by randomly sampling observation times from the population of times available on an *a priori* basis, will have a mean that approaches the true population mean for the response as the sample size increases.

The above assumptions are analogous to those that would be stated for tossing a coin repeatedly for some large number of times throughout some period of time. The "true" distribution of the results would be obtained by observing the total number of tosses. Each toss is considered to be independent of the previous toss. Unbiased estimates of the "true" mean can be obtained by randomly deciding ahead of time which tosses of the coin to observe and record.

Given this framework, it is a logical next step to suggest that an experiment could be designed to test the effects of temperature on the results of repeatedly tossing a single coin in which the following phases were defined:

A_1: the coin is tossed at room temperature for some large number of times.

B_1: the coin is tossed at absolute zero for some large number of times.

A_2: return to baseline conditions.

B_2: return to the conditions of B_1.

Since only one coin has been used, an exact control for order effects cannot be obtained.[2] Thus, a reversal design has been used in which there are actually four treatments: A_1; B_1 given A_1; A_2 given A_1 and B_1; and B_2 given A_1, B_1 and B_2. This means that any cumulative effects of prior conditions are completely confounded with treatments. Nevertheless, a reasonable estimate of the effects of room temperature *versus* absolute zero can be obtained by combining phases A_1 and A_2 and phases B_1 and B_2. This procedure yields two treatments A and B with a large number of independent observations within each. A traditional one-way analysis of variance can test the hypothesis that the two temperatures have different effects upon the coin tossing.

Returning again to the Klein problem, the analogous combined phases for A, B, and C constitute the between-subjects treatments in the one-way analysis of variance, with the observations having been determined by an *a priori* time sampling schedule.

Klein's data in Table 1 provide a vehicle to which the model can be applied. For this purpose, the data for James and Lynn can be used and they are presented in Table 2 (Phases $A_1 + A_2 =$ treatment A; $C_1 + C_2 =$ treatment C). The dependent variable is number of on-task responses.[3] A one-way analysis of variance on these data for James yielded an $F = 3.59$, $df = 2{,}2602$, $p < 0.05$. Thus, the hypothesis that the treatments did not differ in their effect on the on-task behavior of James can be safely rejected. A similar analysis could be done for Lynn's data, but we shall demonstrate that analysis in the two-way model.

Table 2

Mean proportions, standard deviations, and number of observations (recorded intervals) of on-task behavior for each treatment for two subjects from Klein (1971).

		Treatment		
		A	*B*	*C*
James	Mean Proportion	0.274	0.331	0.313
	Standard Deviation	0.446	0.471	0.464
	No. of Observations	1184	610	811
Lynn	Mean Proportion	0.262	0.446	0.364
	Standard Deviation	0.440	0.498	0.482
	No. of Observations	1113	480	708

VIOLATIONS OF THE ASSUMPTION OF INDEPENDENCE OF OBSERVATIONS

Shine and Bower (1971) proposed a one-way model in which they likewise assume "that the subject may be viewed as a response generator the responses of which to a particular stimulus are statistically independent and normally distributed about a central response value." (p. 112) Their model, however, introduces a pseudo-factor, trials, so that the model is actually a two-way design with treatments and trials as independent variables. With this design, there is only one observation per cell which, therefore, requires that the usual within-cell variance estimates cannot be used as the error term for the main effects and interaction as is standard in a fixed-effects model. Although

[2]Typical balanced designs used to control for order effects do not eliminate order effects in any case. All they do is serve to distribute such effects equally across conditions or subjects when data are pooled (see, for example, Sidman, 1960, pp. 245-256). To establish the functional relation between order of treatment and the observed behavior requires, not a balanced design, but a deliberate, systematic manipulation of sequences of treatments, the effects of which are compared with a stable baseline of performance.

[3]Although the dependent variable used here is a dichotomous measure, the F-test is robust with regard to such data (Hsu and Feldt, 1969; Lunney, 1970). Many studies for which this design would be appropriate have available the option of a continuous dependent measure. In Klein's study, for example, this could have been accomplished simply by recording the number of seconds of time on task.

Shine and Bower present a solution to this dilemma, we believe that it is entirely consistent with the assumptions of the fixed-effects analysis-of-variance model to dispense with the trials pseudo-factor *for the reversal design.*

The major objection raised to the application of analysis of variance models to single-subject experiments is that there may not be independence from observation to observation or from treatment to treatment. For example, there may be an observation to observation correlation, in which adjacent observations may be more highly correlated than nonadjacent observations. This dependency could produce problems for studies in which each treatment in succession is applied only once (although it is possible to test this assumption for any set of data). This is not a problem for the reversal design, however, for the reason that observations in adjacent treatments would, by this argument, be expected to be more highly correlated than observations in nonadjacent treatments. Thus, the correlation between observation 1 in Treatment A_1 and observation 1 in Treatment B_1 would be expected to be higher than the correlation between observation 1 in Treatment A_1 and observation 1 in Treatment A_2. Since we combine treatments A_1 with A_2 and B_1 with B_2 for the F-test, then any such correlations between observations will tend to make the treatments more similar and, therefore, reduce the size of the F-statistic. The effect of nonindependence of observations for the reversal design, in short, is to operate in the conservation direction for the F-test.[4]

THE TWO-WAY ANALYSIS-OF-VARIANCE MODEL

It is a straightforward extension of the one-way model to the two-way fixed-effects model. Treatments remain as one factor and subjects become the other. With additional subjects, each one considered a different level of a factor, subject differences can be assessed, as well as subject by treatment interaction effects. Klein's data in Table 2 again provide a vehicle for applying the model, much as we would have had we used two coins in the hypothetical temperature-coin flipping experiment described earlier.

Table 3

Analysis of Variance Summary Table

Source of Variation	*df*	*MS*	*F*	*p-level*
Subjects	1	1.2345	5.80	<0.02
Treatments	2	5.4381	25.55	<0.0001
Subjects X Treatments	2	1.5645	7.35	<0.0007
Within Cell (Error)	4900	0.2128		
Total	4905			

NOTE: This analysis was computed by Finn's (1967) Multivariance program, which accounts for unequal Ns through the least-squares method.

Table 3 presents the traditional two-way analysis-of-variance summary table for the data in Table 2. The between-subjects main effect yielded an $F = 5.80$, $df = 1,4900$, $p < 0.02$. This indicates that Lynn performed significantly more on-task responses than James across all treatments. The between-treatments main effect gave an $F = 25.55$, $df = 2,4900$, $p < 0.0001$, indicating that there were reliable differences among treatments. More interesting, perhaps, is the interaction effect, which yielded an $F = 7.35$, $df = 2,4900$, $p < 0.0007$. This finding provides statistical confirmation for the visual interpretation made earlier that Lynn and James were affected differently by the treatments. Thus, the apparent modest effects of the treatments can now be seen to be statistically reliable.

DISCUSSION

It should be noted that statistical confirmation that a significant difference was obtained in no

[4]In this regard, inclusion of Klein's treatment B in the analysis is not strictly appropriate, since it was presented only once. Consideration of the effects of this minor violation are beyond the scope of this paper since these data are included only as a vehicle for presenting the model.

way guarantees the psychological importance of the findings *vis-a-vis* the reversal design any more than it does in any other kind of investigation. If larger effects than those obtained are necessary to convince teachers, parents, clinicians, or experimenters that they should spend the considerable extra time to apply these procedures, then statistical significance will not constitute sufficient proof to be convincing. Rejection of the null hypothesis does, however, encourage the experimenter to continue to refine the technique in the high probability that he is not wasting his time and efforts on a nonexistent effect.

With that as a general caveat, let us turn to some specific points that might be raised about this approach. First, although the study used here as an example had widely disparate Ns, it may be (as one reviewer stated) that "one should be quite scrupulous to avoid the problem of unequal number of observations per cell." This problem is not unique to this design, of course, and should be considered in the design of any experimental study (*e.g.*, see McNemar, 1969, pp. 118-121). A solution to the problem in the type of study under consideration here would be the use of a continuous dependent measure (as suggested in Footnote 3), collected at pre-planned intervals of equal lengths for each treatment.

A second point has to do with the generalizability of results in a fixed-effects model. Any single-subject or small-N study that shows significant treatment effects should be interpreted as indicating that, for the particular subject or subjects studied, the variance attributable to treatments was sufficiently larger than one might expect by chance. Generalization to other subjects must, in any case, be demonstrated by further study and not merely assumed.

Third, the data from any experimental study may be treated in many ways, and the model we propose is not exceptional in this regard. Thus, with only two treatments and one subject, it may be more appropriate to use a t-test analysis. Where order effects of treatments can be randomized, which is seldom the case in behavior analysis studies, it may be more appropriate to use Latin square arrangements. Or, as one reviewer suggested, it would be possible to consider the A_1 B_1 A_2 B_2 design for one subject (our one-way ANOVA Model) as being a two-way design: Treatments (A *vs.* B) arranged independently of Times (First *vs.* Second). For two or more subjects (our Two-Way ANOVA Model), the classification would then become three-way. In either case, one could obtain separate estimates of sequence and treatment effects, as well as their interaction.

However conceptualized, it seems to us that the analysis-of-variance models proposed here can aid in the interpretation of experimental treatment effects.

REFERENCES

Bandura, A. *Principles of Behavior Modification,* New York: Holt, Rinehart and Winston, Inc., 1969.

Finn, J. D. "Multivariance-Univariate and Multivariate Analysis of Variance and Covariance: A FORTRAN IV Program," unpublished manuscript. State University of New York at Buffalo, 1967.

Hsu, T. and Feldt, L. S. "The Effect of Limitations on the Number of Criterion Score Values on the Significance Level of the *F*-Test," *American Educational Research Journal,* 1969, **6**, 515-527.

Klein, R. D. "The Effects of a Systematic Manipulation of Contingencies Upon Overt Work Behavior in a Primary Classroom," unpublished Ph. D. dissertation, Department of Educational Psychology, State University of New York at Buffalo, 1971.

Lunney, G. H. "Using Analysis of Variance with a Dichotomous Dependent Variable: An Empirical Study," *Journal of Educational Measurement,* 1970, **7**, 263-269.

McNemar, Q. *Psychological Statistics,* New York: John Wiley & Sons, Fourth Edition, 1969.

Shine, L. C. II and Bower, S. M. "A One-Way Analysis of Variance for Single-Subject Designs," *Educational and Psychological Measurement,* 1971, **31**, 105-113.

Sidman, M. *Tactics of Scientific Research,* New York: Basic Books, 1960.

Received 29 April 1971.
(Revised 3 December 1971.)