ASSESSING THE EFFECTIVENESS OF PROGRAMMED GENERALIZATION¹

JOHN D. CONE

WEST VIRGINIA UNIVERSITY

Issues related to assessing change and retention of change were discussed. An alternative analysis was suggested for the data of a recent study by Walker and Buckley (1972). These authors had found that peer reprogramming, equating stimulus conditions, teacher training, and control groups maintained 77, 74, 69, and 67%, respectively, of appropriate behavior produced in a token economy. Their analysis made no use of baseline levels. Two analyses incorporating baseline scores were suggested. One involved change scores; the other, analysis of covariance using baselines as the covariate. Problems with the data made a clear preference difficult, but it was concluded that either analysis would have resulted in conclusions different from those of Walker and Buckley.

It is incumbent upon applied behavior analysts to make obvious the generality of behavior changes they produce (Baer, Wolf, and Risley, 1968). O'Leary and Drabman (1971) reviewed reports of classroom applications of behavior analysis and found little evidence of generalization from token economies to regular classes. Recently, however, Walker and Buckley (1972) demonstrated such generalization. Moreover, these authors found some generalization strategies to be more effective than others. Thus, whereas "overall, subjects maintained 72% of the average amount of appropriate behavior produced in the experimental classroom" (p. 219), comparable values were 77, 74, 69, and 67% for peer reprogramming, equating stimulus conditions, teacher training, and control groups, respectively. Walker and Buckley concluded that "the teacher training maintenance strategy was much less effective than either the peer reprogramming or the equating stimulus conditions strategy in the

amount of appropriate behavior produced during maintenance . . ." (p. 220).

Careful consideration of the Walker and Buckley paper is important for two reasons: (a) their conclusions have serious implications for the application of behavior analysis in the classroom, especially for teacher training, and (b) their data analysis raises a number of general issues related to assessing change and the generalization of change over time and setting. The present paper draws attention to these issues using the Walker-Buckley data to illustrate certain points.

According to the authors, the purpose of the study was to compare ways of "facilitating generalization and maintenance of *treatment effects* after two months in a token economy classroom" (p. 209, emphasis added). Thus, children were randomly assigned to one of the four conditions at the end of treatment and observations continued in regular classes for two additional months. Per cents of appropriate behavior were presented for baseline (regular class), treatment (token economy), and follow-up (regular class) phases.

There are at least two ways of defining "treatment effects" (Lindquist, 1953, p. 1): (a) in terms of within-group or within-subject change over time, and (b) in terms of between-group

¹The author is indebted to Hayne W. Reese for helpful comments on an earlier draft of this paper, and to the anonymous reviewers whose suggestions have greatly enhanced its quality. Reprints may be obtained from the author, Department of Psychology, West Virginia University, Morgantown, West Virginia 26506.

differences, usually assessed at one point in time. The tradition in applied behavior analysis has favored measures involving change. However, the situation is complicated when one moves from demonstrations of the efficacy of single treatments to comparisons among several. An applied analysis still requires that behavior *change* be evaluated. The problem is to find designs that incorporate change and at the same time allow for comparisons among treatments.

Walker and Buckley have not emphasized change in their evaluation of generalization. Rather, they focused on differences between groups in treatment and follow-up phases, with separate analyses being performed for each phase. Before considering the adequacy of their analysis, an alternative procedure using baseline to treatment change scores will be evaluated. The retention of these changes during follow-up becomes the major dependent variable for assessing generalization.

Overall, subjects had a baseline mean (B) of 44.6% appropriate behavior. Subtracting B from the mean of the treatment phase (90.3%) results in a mean change (C) of 45.7%. Loss (L) during follow-up may be determined by subtracting the follow-up mean from the treatment mean ($90.3 - 65.3 = 25.0\%$). The per cent of change actually retained is provided by $(C - L)/C = [(45.7 - 25.0)/45.7] = 45.3\%$. The comparable values are 44.8, 52.2, 49.0, and 33.0 for peer reprogramming, equating stimulus conditions, teacher training, and control groups, respectively. Thus, teacher training does somewhat better than peer reprogramming and somewhat worse than equating stimulus conditions when looked at in this manner. The reliability of these differences would, of course, need to be tested.

Unfortunately, the use of a baseline to treatment change score as the measure of treatment effects in the Walker and Buckley study would be complicated by the combined effects of incomparable baselines and the ceiling imposed by presenting the data in percentage form. The

90.3% mean appropriate behavior during treatment shows subjects were brought very near the ceiling for this measure. Allowing for unreliability they may actually have been at the ceiling. The ceiling effect could influence the results of the study in several ways. Groups with higher baseline means could gain less in terms of baseline to treatment change. For this reason, the same groups would tend to show greater loss from treatment to follow-up when loss is represented as a percentage of gain.

Related to the ceiling problem is the possible unreliability of change scores (Lord, 1963) and, particularly, ratios based on change scores. However, dependent measures that incorporate change reflect the intent and logic of applied behavior analysis more faithfully than do "one time" or "one phase only" dependent measures, and are to be preferred in generalization studies. The reliability of such measures can be assessed separately for each investigation to determine their suitability. Such an assessment would have to be made of the original Walker-Buckley data before the appropriateness of a change score analysis could be determined.

If change scores were rejected as unsuitable, the type of analysis actually performed by Walker and Buckley might be undertaken. These authors analyzed their data in terms of differences between groups during the treatment and follow-up periods. A one-way analysis of variance showed no significant group differences during treatment. A similar analysis did show significant group differences during follow-up, however, and subsequent *post hoc* comparisons of each treatment group with the controls showed significant differences for peer reprogramming and equating stimulus conditions, but not for teacher training. It is apparently these data that led to the conclusion that teacher training was a much less effective strategy. However, such a conclusion would require comparisons between the teacher training condition and the peer reprogramming and equating stimulus conditions, which were not reported. Thus, even if one were to accept the definition of treatment

effects adopted by Walker and Buckley he would still not be able to accept their conclusions without these additional comparisons.²

Further problems posed by the data suggest conclusions different from those of the authors. The baseline means for each of the groups were substantially different, ranging from 36.9 to 53.9%. Consider the fact that, from their unequal baselines, all groups were brought to approximately equal means during treatment. Moreover, these means were at (or very near) the maximum attainable. Treatment lasted two months for all groups. With this information it is possible to predict follow-up means from baseline data alone. The prediction is based on the fact that future probabilities of behavior are linearly related to the number of their previous reinforced occurrences. Thus, subjects entering treatment with higher operant levels of appropriate behavior were likely to be consequted more frequently during the two-month period than were subjects entering with lower operant levels. As a result, appropriate behavior of the former subjects should have extinguished more slowly than that of the latter. These subjects should then have retained more at follow-up even if all maintenance strategies were equally effective. Indeed, the Walker and Buckley data show that, disregarding the controls, the treatment groups were identically rank ordered on baselines and on the per cent of treatment behavior maintained during follow-up.

Because of the plausible relationship between baseline and generalization scores, an analysis controlling for baseline differences would have been more appropriate. An alternative to the change score analysis, already suggested, is provided by an analysis of covariance, with baselines as the concomitant variate. Covariance analysis of follow-up scores would probably result in teacher training and equating stimulus conditions groups showing greater generaliza-

tion than peer reprogramming and control groups, contrary to Walker and Buckley's conclusions. However, analysis of covariance is appropriate when random assignment of subjects has been made and the resulting groups have comparable means on the covariate. When the means are not comparable it is difficult to interpret the results of the analysis. For example, it is possible that floor and ceiling effects, rather than treatment effects, could account for group differences in generalization.

There is a further complication in using analysis of covariance with the Walker-Buckley data. Not only were the group means different during baseline, but their variances were also. An F-MAX test on equating stimulus conditions *versus* teacher training groups yielded a ratio of $(19.94)^2/(7.25)^2 = 7.56$, which is significant at the 0.01 level. Such a great disparity in variances, as well as the further possibility of heterogeneous regression coefficients argues for the use of nonparametric covariance analysis (Siegel, 1956), unless a reanalysis of the original data were to uncover correctable errors accounting for the differences.

It should be noted that neither the change score analysis suggested earlier, nor the analysis of covariance with baselines as the concomitant, can be clearly preferred in the absence of a thorough reanalysis of the original data. In either case, however, it would appear that the conclusions would be different from those of Walker and Buckley.

One additional informal analysis made by the authors should be examined. They stated that "there was an inverse relationship between the amount of staff time invested in the three experimental maintenance groups and the amount of behavior maintenance achieved" (p. 221). Unfortunately, the data supporting this conclusion are unclear. Walker and Buckley stated that peer reprogramming, equating stimulus conditions, and teacher training involved 4, 7, and 9 hr of staff time, respectively. However, the description of the procedures used with the peer reprogramming group indicated that each

²Incidentally, these comparisons were provided in section four of the final report of U.S.O.E. Contract OEG 4-6-061308-0571, and they failed to reach significance.

experimental subject in this condition earned points for "himself and his classmates during two, 30-minute periods scheduled twice a week" (p. 213). These reinforcement periods were managed by a staff member during the two-month follow-up period. Eight weeks of twice-weekly half-hour sessions adds to 8 hr of staff time, not four. Thus, the groups appear to differ by 2 hr at the maximum. Moreover, the 1-hr difference between teacher training and peer-reprogramming strategies scarcely supports the economic superiority of the latter, especially given the possible unreliability of this difference.

Walker and Buckley should be commended for providing data on a question of utmost importance to applied behavior analysts. Though it is difficult to agree with their conclusions, discussion of the issues they have provoked should lead to a better understanding of the pitfalls and difficulties accompanying research in the extremely complex area of programmed generalization.

REFERENCES

- Baer, D. M., Wolf, M. M., and Risley, T. R. Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, 1968, 1, 91-97.
- Lindquist, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.
- Lord, F. M. Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change*. Madison, Wisconsin: The University of Wisconsin Press, 1963. Pp. 21-38.
- O'Leary, K. D. and Drabman, R. Token reinforcement programs in the classroom: a review. *Psychological Bulletin*, 1971, 75, 379-398.
- Siegel, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.
- Walker, H. M. and Buckley, N. K. Programming generalization and maintenance of treatment effects across time and across settings. *Journal of Applied Behavior Analysis*, 1972, 5, 209-224.

Received 3 November 1972.

(Revision requested 18 April 1973.)

(Final acceptance 25 July 1973.)

COMMENTS BY REVIEWER A

Cone: Assessing the Effectiveness of Programmed Generalization

The arguments raised by Cone about Walker and Buckley's paper are indeed important and have major consequences regarding Walker and Buckley's conclusions about the maintenance of behavior changes. As Cone points out, one cannot assert that Peer Reprogramming (PR) and Equating Stimulus Conditions (ESC) are more effective than Teacher Training (TT) merely because the former two are significantly different from the control group (C), but the latter is not. Cone is perhaps too kind in not pointing out that Walker and Buckley make many, loose, quasi-statistical statements (e.g., about "reliable maintenance effect", "ratio of maintenance", etc.).

The major thrust of Cone's paper deals with the appropriate method for analyzing Walker and Buckley's data. Cone notes that there were (probably) substantial differences for baseline operant levels among the four groups, and these might well be related to the maintenance operant levels during the follow up period. Also, the token economy operant level is probably useless as a concomitant variable because of ceiling effects. Unfortunately, Cone does not take a firm stand on how to use the baseline operants. It would seem that change scores or analysis of covariance are equally acceptable to him, but this matter is not pursued further.

It seems to me that the matter is really more complicated than indicated by Cone, and I would like to give my impression of the problems. First there are two readily available concomitant variables (or covariates): baseline