DOCUMENT RESUME

ED 138 639 TM 006 277

AUTHOR Cross, Lawrence H.; Lane, Carolyn E.
TITLE Strategies for Analyzing Data from Intact Groups.
PUB DATE [Apr 77]
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Achievement Gains; *Analysis of Covariance; *Analysis of Variance; Control Groups; Educational Alternatives; *Groups; Post Testing; Pretesting; Program Evaluation; Raw Scores; Reading Achievement; Standard Error of Measurement; *Statistical Analysis; Tests of Significance; True Scores
IDENTIFIERS *Nonequivalent Groups

ABSTRACT

Action research often necessitates the use of intact groups for the comparison of educational treatments or programs. This paper considers several analytical methods that might be used for such situations when pretest scores indicate that these intact groups differ significantly initially. The methods considered include gain score analysis of variance (ANOVA), analysis of covariance (ANCOVA) (using both raw scores and estimated true scores), value-added analysis, and within group dependent t-tests, all on a common set of real data from nonequivalent intact groups. Seemingly contradictory results were obtained for this data with gain score ANOVA and with ANCOVA. Comparable results should be expected to occur routinely with data from nonequivalent groups. In view of these results, it is recommended that statistical comparisons across nonequivalent groups be avoided. However, within group comparisons may aid somewhat in such evaluations of alternative educational programs. (Author/RC)

Strategies for Analyzing Data from Intact Groups

Lawrence H. Cross and Carolyn E. Lane

Virginia Polytechnic Institute and State University

Paper presented at the Annual Convention of the American Educational Research Association, New York, New York, April, 1977, Session 8.02.

ABSTRACT

Strategies for Analyzing Data from Intact Groups

Lawrence H. Cross and Carolyn E. Lane

Virginia Polytechnic Institute and State University

"Action research" often necessitates the use of intact groups for the comparison of educational treatments or programs. The purpose of this paper is to consider several analytical methods that might be used for such situations when pretest scores indicate that these intact groups differ significantly initially.

The methods considered include gain score ANOVA, ANCOVA (using both raw scores and estimated true scores), value-added analysis, and within group dependent t-tests, all on a common set of real data from nonequivalent intact groups. Seemingly contradictory results were obtained for this data with gain score ANOVA and with ANCOVA. Comparable results should be expected to occur routinely with data from nonequivalent groups.

In view of these results, it is recommended that statistical comparisons across nonequivalent groups be avoided. However, within group comparisons may aid somewhat in such evaluations of alternative educational programs.

Strategies for Analyzing Data from Intact Groups

Lawrence H. Cross
Carolyn E. Lane

Virginia Polytechnic Institute and State University

The term "action research" suggests research which is responsive to the immediate needs of a decision maker in a particular setting. As such, time and administrative obstacles may argue for the use of extant groups to compare two or more treatments or programs. Since random assignment of subjects to groups is not possible, such a research design would be considered a quasi-experimental, non-equivalent control group design in the Campbell and Stanley (1963) taxonomy. With such designs, it is advisable to pretest the subjects to determine the extent to which the intact groups differ with respect to the variables under study. If the mean pretest scores for the groups do not differ significantly, one may wish to assume that the groups were, in effect, randomly formed and proceed with an analysis appropriate for a true experiment, including any of these which follow.[1] If, however, the groups differ significantly on the pretest, indicating the groups are not likely to represent random samples from a common population, there is little agreement regarding how such data should be treated. The purpose of this paper is to consider a number of analytic methods that might be used with such data. In order to facilitate comparisons, each analysis reported

[1]Note that failure to reject a null hypothesis does not imply the truth of the null. Moreover, the groups may differ considerably with respect to some unmeasured but relevant variable. Consequently, this strategy is a poor substitute for random assignments to groups. At issue is whether the groups can be considered equivalent or non-equivalent in both a statistical and a practical sense.

below was carried out using a common set of real data.[2] Pre- and posttest scores from the reading subtests of the Metropolitan Achievement Tests, Primary Level II, were obtained for children in each of three intact groups. Each group was instructed using a different reading program during the course of the academic year. The tests were administered in the fall and the spring to the children in all groups. Pretest scores were not available when the groups were formed. Even though the Metropolitan provides three subtest scores for reading (word knowledge, word analysis, and reading), only the total reading scores were used in the analysis reported here. The total reading scores are obtained by summing the number of correct responses across the three subtests. Ordinarily, a multivariate analysis of the subtest scores would be preferred, but univariate analyses using total reading scores are reported in this paper to facilitate the discussion. In practice, the parsimony achieved by the use of a single composite score is gained at the expense of diagnostic information that a multivariate analysis of the subtests would have afforded.

ANOVA on Gain Scores

Perhaps one of the most obvious analyses for data of this type is to compare the raw gain scores across groups. While it is true that gain scores tend to be highly unreliable, this characteristic of gain scores is of greatest concern when gain scores are to be used in a correlational study. The

[2]The writers wish to express their appreciation to Dr. Rose Sabaroff for providing us the data for the analyses reported in this paper.

unreliability of gain scores has been shown not to be a valid concern w n the interest is to compare differences between experimental treatment groups (Overall and Woodword, 1975).

Table 1 presents the means and standard deviations of the pretest, post-test, and gain scores earned by the three groups. An analysis of variance using the gain scores revealed that the differences are significant ($p < .001$) and a Newman-Keuls post-hoc test indicated that the pairwise differences among all three groups were also significant ($p \leq .01$). If one were to present results such as these to a decision maker who is not well versed in the ways of gain scores, he might well decide against the programs used with groups I and III and choose the program used with group II. You might feel obliged, as an action researcher, to explain that the smaller gains observed for group I may simply reflect the fact that the group was of higher ability to begin with and there was less room for improvement in this particular test in comparison to the other groups.[3] Thus, had the groups been of equal ability at the beginning of the year, the analysis of gain scores may lead to quite a different conclusion.

Analysis of Covariance

Rather than attempt to explain to a decision maker that "what you see is not what you get," due to pre-existing differences between groups, you may decide to use the analysis of covariance (ANCOVA) to ". . . make adjustments

[3] If the groups had been formed on the basis of the pretest scores, the regression toward the mean phenomenon might also be used to explain such a result. Such was not the case with these data.

for the effects of the uncontrolled variables in comparing group performance" (Tatsuoka, 1971, p. 40). In this example, ANCOVA was used to "control" for pre-existing differences in reading ability as measured by the pretest. Essentially, the analysis of covariance adjusts the group's mean scores on the dependent variable(s) or posttest scores as a function of the group's performance on the covariate. The slope of the regression line of the posttest on the pretest is used to make the "appropriate" adjustment, and it must be assumed that the slope of the regression does not differ significantly across groups. (A conservative level of α should be used in this test since the objective is to show that the null is tenable.)

When the ANCOVA was applied to the reading scores, the assumption indicated above was well satisfied ($p \leq .90$) and the differences among the adjusted posttest means were found to be significant ($p \leq .002$). A consideration of the adjusted posttest mean scores, which are also presented in Table 1, suggests that, after initial differences in ability are adjusted, the reading program used with group III was not nearly as effective as those used with groups I and II. Before attempting to convince the decision maker that the results of the ANCOVA are to be believed over the ANOVA on gain scores, one might wish to consider a modified ANCOVA.

ANCOVA Using Estimated True Scores

Lord (1963) has pointed out that "making allowances for initial differences among groups on a poor measure of some variable is not the same thing as making allowances for initial differences in the variable itself." The procedure suggested by Lord (1960) to overcome this problem requires

the administration of the same pretest twice in order to arrive at the estimated true scores. Since even in the best of situations, it would be rare to be able to administer two pretests to all subjects in all groups, a "reasonable" alternative to this was taken in order to obtain estimated true scores on the pretest for the data reported herein.

By using the classical measurement assumption that the standard error of measurement is constant over the ability range measured by a test, it was possible to estimate the reliability of the test in this setting by substituting the standard error of measurement provided by the test manual in the following formula:

$$s_{meas} = s_x\sqrt{1 - r_{xx}},$$

substituting the pooled posttest standard deviations of the pretest scores for s_x and solving for r_{xx}. Using this estimate of r_{xx}, the estimated true scores were computed using:

$$T = X + r_{xx}\,(X - \bar{X}_g),$$

where T is the estimated true score, X is the observed score and $\bar{X}_g$ is the mean of the group to which each subject belongs. In words, each person's score was regressed toward the mean of his group as a function of the estimated reliability. When the estimated true scores so determined were entered into the usual analysis of covariance procedures, slight differences were observed in the adjusted posttest scores as shown in Table 1. In this application, the effect was small since the composite total reading scores were already highly reliable. With less reliable covariates, however, the use of estimated true scores may substantially alter the results of the ANCOVA (Lord, 1960).

The use of ANCOVA using estimated true scores was included here since it is recommended by Porter and Chibucos (1974) as the preferred method of analysis with data from non-equivalent control group designs.

A Comparison of Gain Score ANOVA vs ANCOVA

The fact that the gain score analysis and the analyses of covariance reported above give seemingly contradictory results is not an artifact of these particular data, but can be expected to occur routinely with data from non-equivalent control group designs. The analysis of covariance simply anticipates and adjusts scores so as to account for the phenomenon referred to as regression toward the mean. When any group is measured twice on the same variable, there will be a tendency for the high or low scoring individuals (or subgroups) to regress toward the mean unless everyone earned the same score on both occasions.[4] Note that a person's score is regressed toward the mean of the group of which he is a member or can be assumed to be a member. It does not make sense to regress a person's score toward the mean of a group if he could not reasonably be assumed to belong to the group. The latter, however, is essentially what the analysis of covariance does when it is applied to data like that reported here. Only if the pretest means do not differ significantly is it reasonable to regress these means toward a common population mean. Lord (1967) offers a vivid illustration of the perils associated with using ANCOVA when a single treatment is applied to samples drawn from two distinct populations. The point made by Lord can be illustrated

[4] Low reliability may contribute to the regression effect but even if perfectly reliable measurements are taken, the regression effect should be anticipated as long as the correlation between pre- and post- scores is less than perfect.

with the present example by considering what would happen if the pretest had been given in May and the posttest had been given the following October. In such a situation, some pupils would be expected to gain over the summer and others would lose, but it might be reasonable that by October, the pretest and posttest means would be nearly the same within each ability group. Such an outcome was approximated with the present data by subtracting from each person's posttest score an amount equal to the difference between the pretest and posttest means for his group. Making the pretest and posttest mean scores equal within each group, while the individual scores are free to change, represents a condition Lord refers to as dynamic equilibrium. An analysis of covariance was then applied with the result that, after "controlling" for the pretest differences, the groups were found to differ significantly ($p < .001$). The adjusted posttest means are shown in Table 1. Inasmuch as no group gained or lost, it may be a bit awkward trying to explain to the decision maker how the summer had a significantly more favorable effect on the high ability group in comparison to with average and low ability groups. Notice that each group was exposed to the same treatment, summer. When each group is exposed to a different treatment, the explanation becomes even more tedious, if not absurd.

Studies which provide data of the type reported here may have been designed as single factor studies, but, by default, become two factor studies when the groups are found to differ significantly prior to treatment. It is impossible to disentangle the effects of the two factors unless each treatment is applied to each ability group. Moreover, the analysis of covariance cannot eliminate the confounding of the ability factor by making equal that which God made unequal. It is for these reasons that the writers recommend against the use

of ANCOVA to analyze data of this type. Such advice is consistent with that offered by some (e.g., Elashloff, 1969; Cronbach and Furby, 1970; Lord, 1963, 1967), but is counter to the advice offered by others (Campbell and Stanley, 1963, p. 49; Ferguson, 1971, p. 288; Tatsuoka, 1971). Nor will use of estimated true scores resolve the difficulty because the problems indicated above hold even for perfectly reliable measures on the covariate. (Note that in the procedure outlined above for getting the estimated true scores, the observed scores were regressed toward the means of the respective groups, not toward the grand mean.)

The analysis of gain scores should be preferred over analysis of covariance in non-equivalent control group design if only because it is more easily understood and requires fewer assumptions. Once regression toward a common mean is eliminated from consideration, what factors, if any, argue against a straight-forward interpretation of gain scores? One factor which seems to have been operative in the present study was an artificial ceiling effect associated with the use of this particular test. This effect is evident by the fact that the mean posttest scores for the high ability group ($\bar{X}$ = 105.5) was close to the maximum possible score (119). Were it not for this ceiling effect, it might be reasonable to expect the high ability group to maintain or increase their superiority by gaining the most. While this effect works in the opposite direction of the ceiling effect, it is not reasonable to assume the two will balance each other. The only interpretation that can be drawn from a gain score ANOVA as reported here is that the amount of gain was significantly different when program I was used with "high" ability pupils, program II was used with "average" ability and program III was used with "low" ability pupils.

Alternative Analyses

While it does not seem reasonable to make comparisons across treatments, it may be of interest to consider within treatment comparisons. For example, it may be of interest to test whether the mean gain observed within a particular program represents a statistically significant gain. The dependent t-test would be appropriate for such a test. Applying the dependent t-test to the data reported here, the t-values were all highly significant. The inference to be made from these tests is with reference to subsequent samples drawn at random s mple from the three distinct ability populations associated with each group. In this application, the gains within all three groups were so large that a statistical test of the null hypothesis may seem trivial. It may, however, be of interest to test whether the observed gain is significantly different from some *a priori* expectation of gain based on practical or theoretical considerations, rather than to test against a null hypothesis of zero gain. One of the more interesting proposals in this regard is that by Bryk and Weisberg (1976) called Value-Added Analysis. Very briefly, the pre- and posttests are viewed as snapshots of an on-going developmental process and chronological age (or some other variable) is regressed onto the pretest scores to provide an estimate of the growth that might be expected without special intervention. Unfortunately, when this strategy was applied to the data reported here, it was found that the regression of age on pretest scores was nearly zero which argued against the use of this new analysis.[5]

In certain situations, it may be of interest simply to determine whether the observed mean gain could be attributed to errors in measurement alone. For example, if a test had been administered to this audience before and again after this presentation, it might be of interest to determine whether

[5] This outcome was quite disappointing since it seemed reasonable to find some relationship between chronolog al age and reading ability for children in "regular" third grade classes.

the observed gain (regardless of sign!) represents a difference larger than what might be attributable to error. The standard error associated with sampling error t in this case since we do ish to generalize beyond this udience. The procedures outli Davis (1964) to estimate the standard error of measurement of the mean change should be used rather than the usual dependent t-test. While the standard error of mean change is conceptually distinct from the standard error of measurement of mean change, operationally the distinction can be lost. Specifically, the variance of the difference scores used in the dependent t-test can also be taken as an estimate of the variance error of measurement if an additive treatment model can be assumed (Rulon, 1941; Overall and Woodword, 1975). Viewed in this way, it seems inappropriate to use the results of a dependent t-test to make inferences regarding subsequent samples drawn from the same population since the standard error would reflect measurement error and not sampling error.

Summary

If it is necessary to go beyond simple descriptive statistics resulting from non-equivalent group designs, statistical comparisons across groups should be avoided. Within groups comparisons are logically consistent, but whether the results should be used to make a statistical inference or a measurement inference must be considered. If these recommendations do not offer much help for analyzing data of this type, so be it. Perhaps it is time for action researchers to educate decision makers regarding the importance of random assignments to groups if statistical tests are to aid in

the evaluation of alternative programs. If all else fails, the descriptive statistics can be scrutinized carefully and can provide a basis for judgment in much the same manner people choose spouses. If it later turns out that the decision was in error, at least it wasn't the fault of the action researcher who inappropriately used the analysis of variance.

TABLE I

Summary Statistics of Non-equivalent Group Data

	Pretest		Posttest					
	$\bar{x}$		$\bar{y}$	s_y	Raw Gain	Adjusted $\bar{y}$*	Adjusted $\bar{y}$**	Adjusted $\bar{y}$ ***
Group I n = 46	92.		105.52	11.52	12.81	94.09	93.74	81.29
Group II n = 60	62.52	23.57	91.00	17.07	28.48	96.08	96.24	67.60
Group III n = 19	50.53	18.28	74.11	18.79	23.58	85.74	86.10	62.17

*Based on raw score ANCOVA

**Based on estimated true score ANCOVA

***Based on ANCOVA with equal pre- and posttest means within groups

REFERENCES

Bryk, A. S. and Weisberg, H. I. Value-Added Analysis: A dynamic approach to the estimation of treatment effects. *Journal of Educational Statistics*, 1976, *1*, 127-155.

Campbell, D. T. and Stanley, J. C. *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally and Company, 1963.

Cronbach, L. J. and Furby, L. How we should measure "change" -- or should *Psychological Bulletin*, 1970, 7[illegible] 68-80.

Davis, F. B. *Educational Measurements and Their Interpretation.* Belmont, California: Wadsworth Publishing Company, Inc., 1964.

Elashoff, J. D. Analysis of Covariance: A delicate instrument. *American Educational Research Journal*, 1969, *6*, 383-401.

Ferguson, G. A. *Statistical analysis in psychology and education.* New York: McGraw-Hill, 1971.

Lord, F. M. Large Sample Covariance Analysis When the Control Variable Is Fallible. *American Statistical Association Journal*, 1960, [illegible]5, 307-321.

Lord, F. M. Elementary Models for Measuring Change. In C. W. Harris (ed.), *Problems in Measuring Change*. Madison Wisconsin: The University of Wisconsin Press [illegible].

Lord, F. M. A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 1967, *68*, 304-305.

Metropolitan Achievement Tests, Primary Level [illegible]. Harcourt Brace Jovanovich, 1971.

Overall, J. E. and Woodward, J. A. Unreliability of difference scores: A paradox of measurement of change. *Psychological Bulletin*, 1975, *82*, 85-86.

Porter, A. C. and Chibucos, T. R. Selecing analysis strategies. In G. D. Borich (Ed.) *Evaluating educational programs and products.* Englewood Cliffs, New Jersey: Educational Technology Publications, 1974.

Rulon, P. J. A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*, 1939, *9*, 99-103.

Tatsuoka, M. M. *Multivariate Analysis.* New York: John Wiley and Sons, 1971.