DOCUMENT RESUME

ED 137 399 TM 006 205

AUTHOR Raffeld, Paul; Reynolds, William M.

TITLE Increasing the Efficiency of Pretest-Posttest Designs.

PUB DATE [Apr 77]

NOTE 9p.; 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 Control Groups; Correlation; Experimental Groups; Hypothesis Testing; Measurement Techniques; *Post Testing; *Pretesting; *Research Design; Sampling; *Statistical Analysis; Test Reliability; *Tests of Significance

ABSTRACT

The pretest-posttest design referred to as Design 2 by Campbell and Stanley (1963) is commonly used in educational research and evaluation. The tenability of the assumption of a zero population difference commonly used with this design is questioned. A nonzero population estimate based on the mean difference observed in test-retest reliability data is recommended. When a control group is available, it is recommended that the pretest-posttest difference for the control group be subtracted from the experimental group difference. This will produce a more accurate estimate of the magnitude of change for the experimental group. (Author)

ED137399

Increasing the Efficiency of

Pretest-Posttest Designs

Paul C. Raffeld

Office of Evaluation and Development

Louisiana Health and Human Resources

William M. Reynolds

State University of New York at Albany

SESSION # 16.14

Abstract

The pretest-posttest design referred to as Design 2 by Campbell and Stanley (1963) is commonly used in educational research and evaluation. The tenability of the assumption of a zero population difference commonly used with this design is questioned. A non-zero population estimate based on the mean difference observed in test-retest reliability data is recommended, thus allowing for greater control of some of the factors known to affect Design 2 results. When a control group is available, the authors recommend that the pretest-posttest difference for the control group be subtracted from the experimental group difference. This will produce a more accurate estimate of the magnitude of change for the experimental group.

INCREASING THE EFFICIENCY OF

PRETEST-POSTTEST DESIGNS

The one-group pretest-posttest design, referred to by Campbell and Stanley (1963) as Design 2, has been criticized for its lack of validity (Campbell & Stanley, 1963; Kerlinger, 1973). One method of reducing the number of false inferences concerning the existence of a treatment effect when using Design 2, is to test sample differences against a non-zero population estimate.

When statistically comparing two means, the general form of the null hypothesis is $\mu_1 - \mu_2 = k$. In spite of the fact that k can be set to any small value considered of practical interest (Winer, 1971), the overriding tendency is to set k to zero, the expected value, so that the null hypothesis becomes $\mu_1 = \mu_2$. While there is nothing statistically wrong with testing against an expected value, it is obvious that such a practice results in maximizing the sensitivity of the test to any significant difference, regardless of any practical implications.

Although the authors support the more frequent use of values greater than the expected value, this paper addresses the issue of underestimating the expected value and the consequent increase in detecting invalid significant differences. It is in this sense that the word "efficiency" is used in the title rather than in the sense of statistical power. Actually, using a value of k greater than the expected value would decrease the power of the test only if the specific alternative hypothesis used against the zero expected value was maintained, and this would not be likely.

In the case of the t test for independent samples, there is no debate that the expected value is zero, given random sampling and random assignment to treatments. However, it is not reasonable to conclude that an expected value of zero is correct in the case of the t test for correlated samples used in Design 2. It is quite common to find the mean of the second administration of a test to be higher than the mean of the first administration, even over short periods of time and with no deliberate intervention.

Potential estimates of population differences are available in many instances, but they are seldom used. One major reason for this is a concern over the accuracy of such estimates. It is the author's contention, however, that in most instances of pre and post testing, the expected differences will not be zero, and that any estimate of k that is greater than zero, but not in excess of a practical difference should be used. When in doubt, it seems better to overestimate rather than underestimate the expected value.

One estimate of a population difference useful in Design 2 studies can be obtained by examining the test - retest means from test - retest reliability data. This mean difference is usually ignored because the reliability coefficients are typically high. It is easy to forget that the Pearson correlation used to obtain test - retest reliability is insensitive to changes in the mean. Stability over time in the test - retest sense refers primarily to the preservation of relative rank order and does not reflect a change in the mean over time. It is not unusual to have a test - retest reliability coefficient of .90 and a difference between the test - retest means that is statistically significant. The authors have found that the typical mean posttest increase shown in technical reports for IQ tests and achievement batteries is often significant beyond the .05 level. These measures are usually taken over a 2- to -6 week period with no intentional treatment intervention.

As noted earlier, the reluctance to use such estimates of the expected difference reflects concerns related to sampling stability and the comparability of populations. However, these concerns are important regardless of the chosen value of k. It is obviously inappropriate to administer a test designed and normed on one population to a sample from another population and expect to have comparable results. Furthermore, population estimates based upon small heterogeneous samples are not as stable as one would desire. Keeping these factors in mind, however, the authors still recommend using a statistically significant test - retest mean difference as the expected value of k, rather than zero, because the chances are that even this value will underestimate the expected value more often than is desirable.

It seems reasonable to assume that such factors as history, maturation, the effect of pretest administration and statistical regression are reflected to some extent in these test - retest mean differences (Campbell & Stanley, 1963). The relatively short test - retest intervals used in reliability analysis will tend to increase the effect of some of these factors while decreasing the effects of the others. By using the test - retest mean difference as suggested here, the effects of these confounding variables should be reduced and, therefore, increase the validity when rejecting the null hypothesis.

The form of the t test for correlated samples is:

$$t = \frac{\bar{D} - \mu_{\bar{D}}}{S_{\bar{D}}/\sqrt{N}}$$

where $\mu_{\bar{D}}$ is the expected value and $\bar{D}$ represents the sample mean difference. The standard error is based upon the sample data obtained for the study regardless of the value chosen for $\mu_{\bar{D}}$. The recommendation for Design 2 studies then, is to use the test - retest mean difference, if available and significant, as the estimate of $\mu_{\bar{D}}$. When test - retest data are not available, it may still be better to select some arbitrary small value of $\mu_{\bar{D}}$ rather than to use zero.

When a control group is available and Design 2 is unnecessary, the usual approaches to analyzing the data are 1) to compare the raw difference scores for the experimental and control groups using a t test for independent sample or 2) to compare the adjusted posttest means of the experimental and control groups using analyses of covariance with the pretest scores serving as the covariate. In both of these analyses, one is essentially interested in determining whether there is significantly greater gain in the experimental group than in the control group. However, neither of these procedures deal directly with the actual magnitude of the gain. While it is true that in the case of the raw difference score approach the mean difference for each group is available, this difference for the experimental group is quite likely to be greater than it should be as noted earlier. Again, the overestimate is based upon a null hypothesis of $\mu_1 = \mu_2$. Therefore, when control group data is available and a significant difference between the experimental and control groups has been found, it may prove advantageous to use a t test for correlated samples and analyze the gain in the experimental groups using the control group mean difference as $\mu_{\bar{D}}$. If proper sampling and assignment procedures are employed, this analysis should reduce the artificial gain attributed to non-treatment factors. If the accuracy of the estimated control group mean difference is seriously questioned, then one could establish 95 percent confidence limits for the estimate and select the lower bound.

References

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

Kerlinger, F. Foundations of behavioral research (2nd ed.) New York: Holt, Rinehart and Winston, Inc., 1973.

Winer, B. J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill Book Company, 1971, p. 31.