DOCUMENT RESUME

ED 164 571 TM 007 817

AUTHOR Corder-Bolz, Charles R.
TITLE A Monte Carlo Study of Six Models of Change.
INSTITUTION Southwest Educational Development Lab., Austin, Tex.
PUB DATE [78]
NOTE 32p.

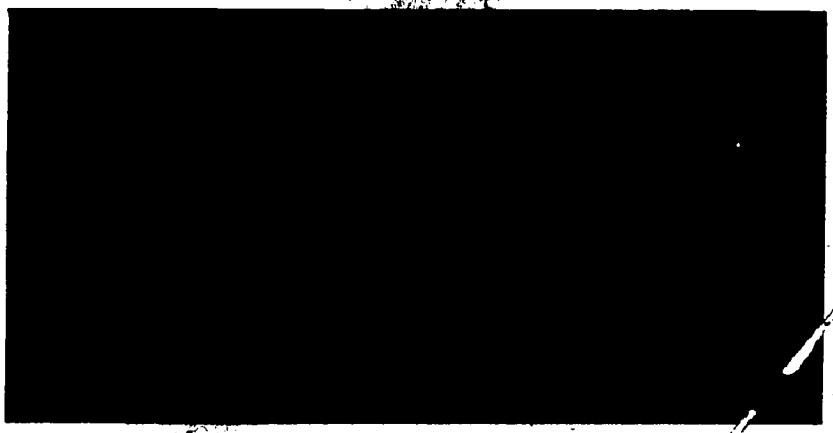
EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS *Analysis of Covariance; *Analysis of Variance; Comparative Statistics; Hypothesis Testing; *Mathematical Models; Scores; Statistical Analysis; *Tests of Significance
IDENTIFIERS *Change Scores; *Monte Carlo Method

ABSTRACT

A Monte Carlo Study was conducted to evaluate six models commonly used to evaluate change. The results revealed specific problems with each. Analysis of covariance and analysis of variance of residualized gain scores appeared to substantially and consistently overestimate the change effects. Multiple factor analysis of variance models utilizing pretest and post-test scores yielded invalidly low F ratios. The analysis of variance of difference scores and the multiple factor analysis of variance using repeated measures were the only models which adequately controlled for pre-treatment differences; however, they appeared to be robust only when the error level is 50% or more. This places serious doubt regarding published findings, and theories based upon change score analyses. When collecting data which have an error level less than 50% (which is true in most situations), a change score analysis is entirely inadvisable until an alternative procedure is developed. (Author/CTM)

RESEARCH SERIES

SOUTHWEST EDUCATIONAL DEVELOPMENT LABORATORY
211 East 7th Street
Austin, Texas 78701



James H. Perry
EXECUTIVE DIRECTOR
Southwest Educational Development Laboratory

For further information contact:

OFFICE OF COMMUNICATIONS
SEDL
211 East 7th Street
Austin, Texas 78701

A MONTE CARLO STUDY OF SIX MODELS OF CHANGE

Charles R. Corder-Bolz, Ph.D.

RESEARCH PAPER NO. 1-02-78

Southwest Educational Development Laboratory
211 East 7th Street
Austin, Texas 78701

A MONTE CARLO STUDY OF SIX MODELS OF CHANGE

Charles R. Corder-Bolz

ABSTRACT

A Monte Carlo Study was conducted to evaluate six models commonly used to evaluate change. The results revealed specific problems with each. Analysis of covariance and analysis of variance of residualized gain scores appeared to substantially and consistently overestimate the change effects. Multiple factor analysis of variance models utilizing pretest and post-test scores yielded invalidly low F ratios. The analysis of variance of difference scores and the multiple factor analysis of variance using repeated measures were the only models which adequately controlled for pre-treatment differences; however, they appeared to be robust only when the error level is 50% or more. This places serious doubt regarding published findings, and theories based upon change score analyses. When an investigator is collecting data which have an error level less than 50% (which is true in most situations), then a change score analysis is entirely inadvisable until an alternative procedure is developed.

A MONTE CARLO STUDY OF SIX MODELS OF CHANGE

Charles R. Corder-Bolz

The desire to observe and understand the forces that cause change is fundamental to educational and social scientists. Change phenomena include such intriguing aspects of life as the acquisition of knowledge, the reduction of anxiety, positive changes in self-concept, and the increase of productivity in human interactions. These phenomena are most validly viewed within the context of change. Therefore, the concept of change is basic to the educational and social science researcher. The measurement of various constructs and their change has reached a high degree of sophistication. The very reliability and validity of such measurements can be estimated. The scientist can choose from a wide array of measurement instruments that include questionnaires, interview techniques, and observation procedures. The critical issue, however, is one of how the scientist can evaluate the observed changes and choose from among various contrasting hypotheses regarding the nature of the change phenomena.

There are two broad categories of methodologies of the study of change. The first category includes the various approaches based upon experimental design considerations. Characteristically, experimental design approaches utilize two or more parallel groups which receive different treatments. The analysis of variance model can then be used to analyze the post-treatment scores. The intent is to assess change

through the observation of differences between groups caused by the various treatments administered to the different groups. Random assignment to the groups should result in independent and equivalent samples. Unfortunately, true randomization is difficult in the "real" world and, thus, there are often important differences between the groups prior to the administration of the treatments. These pre-treatment differences sometimes have profound influences on post-treatment group observations. Consequently, researchers have a serious desire to control initial or potential-initial differences between treatment groups. This desire leads them to the second category of approaches based upon mathematical methods designed to eliminate pre-treatment differences between groups when evaluating changes in those groups.

A number of statistical models and computational procedures are included in this second category. Probably the most commonly used statistical approach to the control of pre-treatment differences is that of difference scores or simple change scores. This approach involves the subtraction of a pre-treatment observation from each post-treatment observation. Thus, a subject's change score is derived by subtracting his pretest score from his post-test score. The result is theoretically the change caused by the treatment. If the measurement tool used has a 100% reliability, then this difference score should be a valid measurement of the change. The concern arises from the fact that most instruments used to measure educational or behavioral change are plagued by a degree of unreliability. Unre-

ability is, in effect, error in the measurement process. It is assumed that this error is independent in each given observation. Therefore, each observed score is a function of what can be called a true score plus the measurement error associated with the observation. If people are tested, whether pretested and/or post-tested, then each test score is composed of both a true score and the independent error component. If the people do not change between the two testings, then the two true scores for each should equal each other. In this case, obtained difference scores would contain no really true score, but rather be composed entirely of error. Likewise, theoretically, in situations where there is change, a difference score would contain the difference between the true scores, or true-score delta, plus the error associated with the first measurement and plus the error associated with the second measurement. Though a measuring instrument may yield data with an acceptable level of error, the difference scores resulting from two uses of the measure could contain a very high level of error. For example, if a questionnaire had a reliability of .90, then approximately 80% of the variance of the scores would be so-called true scores and approximately 20% of the variance would be error. If a treatment increased the true scores by 10% or accounted for 10% additional variance in the true scores in the post-test, then the true-score delta component of the difference score should reflect the 10% variance in the post-test true score that is independent of the pretest true score variance. However, the error associated with the second as well as the first measurement are

independent of each other and independent of the true scores. Therefore, the error component of the difference score would include, or could include, the 20% error variance from the first measurement plus the 20% error variance from the second measurement. The error level of the difference score would likely be the 20% associated with the measuring instrument and could be as high as the sum of the error levels associated with both of the two measurements. Theoretically, such a difference score would have a signal-to-noise ratio of 1:2 and could be as high as 1:4. This is in contrast to the signal-to-noise ratio of 8:2 normally associated with the questionnaire. Therefore, from a measurement theory perspective, the use of difference scores or simple change scores is very questionable.

DuBois (1957), Lord (1956; 1963), and McNemar (1958) have recommended the use of "residual gain" scores as a preferable substitute to the use of "raw gain" of scores. In this procedure, a gain is expressed as the deviation of post-test score from the post-test/pretest regression line. Thus, the part of the post-test information that is linearly predictable from the pretest can be partialed out. The residual, or the residualized gain, is then used to evaluate the change by eliminating any pretest differences or biases in the difference scores. A concern with the use of residualized gain scores is the consequence of partialing out the pretest information. The information that the post-test and pretest scores have in common is what can be considered as the true score component of the pretest score. This true score component of the pretest score

has predictive value. The error component of the pretest score has no predictive value and, therefore, has no function in the regression procedure. Ostensibly, when the pretest is used as a predictor, the effect is to remove the true score information from the post-test scores. The concern is that the residual or that which the pretest cannot predict is, in effect, the error in measurement plus the possible gain in true scores. Consequently, as with difference scores, residualized gain scores run the risk of being primarily composed of error.

A fourth approach to the analysis of change is the multiple factor analysis of variance model. The treatment conditions are represented as a dimension in the analysis and the pretest versus post-test scores are two levels in an additional dimension. The effect is to partition or separate the various sources of variance such as treatment effects and pretest effects. With this model, the investigator is able to isolate and evaluate possible pretest differences among the subjects as well as possible treatment differences between the subjects. If there is a difference or change due to one or more of the treatments, this will result in a greater pretest to post-test difference for one of the treatments in comparison to the other treatments. This effect will be reflected in the interaction component of the analysis. Specifically, the change is evaluated by the F ratio of the mean square interaction over the mean square error. However, the analysis of variance model assumes an independence among all observations. In the present situation, pretest and post-test

measurements cannot be assumed to be entirely independent. Thus, this particular model is rarely used and is included herein mainly for comparison purposes.

A fifth approach to the analysis of change involves a refinement of the analysis of variance model which accommodates multiple measurements derived from the subjects. This procedure, which is commonly referred to as a repeated measures analysis of variance, is computationally similar to the above-mentioned method of multiple factor analysis of variance. However, there are additional sums of squares and mean squares which reflect the effects of between-subject differences and the interaction between the subject and treatments. As with the multiple-factor analysis of variance, change is evaluated by the pre-post test and treatment interaction term. However, despite the d nctions made in the theoretical foundations, the F ratio for this int -ac ion should have exactly the same value as the F ratio generated by a one-way analysis of variance utilizing difference scores (Jennings, 1972).

A xth approach to the analysis of change is the analysis of covariance model. The pretest is used as a covariate in an attempt to control for pretest differences between subjects. The variance of the post-test scores that is linearly predictable from the pretest scores is partialed out. The model is similar to the analysis of variance of residualized gain scores except that the former is based upon within-treatment group regression whereas the latter is based upon a regression across the entire sample. One of the problems with this approach

is that the traditional covariance model assumes an independence of measurement of the covariate and the dependent variable. More specifically, there is a necessary assumption of the independence of the error associated with each of the two measurements. Clearly, the use of pretest scores as a covariate to analyze post-test scores violates this assumption. Furthermore, the analysis of covariance model also theoretically suffers the problem of high error levels. When the pretest is used as a covariate, the result is the removal of the true score information from the post-test scores that is also contained in the pretest scores. The residual is the error of measurement plus any change in the true score values. The resultant information could contain a disproportionate amount of error.

The issues of the evaluation of change remain unresolved because the various theoretical positions approach the problem from different assumptions, and therefore have no common ground from which a common assessment can be made. A particularly important difference in perspectives is the concern over the proportion of error in change scores. The research community was stunned, if not confused, by Overall and Woodward's (1975) demonstration that the power of tests of significance is maximum when the reliability of the difference scores is zero. The best advice to date had been not to measure change at all (Cronbach & Furby, 1970).

In situations in which the uncertainties cannot be resolved in a theoretical manner, insight can often be gained from a Monte Carlo study. In this kind of study, artificial data is generated such that

they conform to a desired structure. Various data sets are generated to reflect the various differences of data parameters that are of concern. Then one or more data analysis models are used to analyze the data to determine the extent to which the models give valid results. In the case of the evaluation of change, insight might be gained by generating sets of data with known characteristics, such as error level, treatment effects, and pretest differences, then applying the various approaches to the data sets. The results would provide a basis for a direct comparison of the models.

Method

The basic method was to simulate the traditional treatment versus control group experiment in which each subject is pretested and post-tested. The two groups were composed of randomly assigned subjects, with an arbitrary number of 20 subjects per group. One group represented the treatment group which received some kind of experimental treatment and the other group represented the traditional control group which either received no treatment or a neutral treatment. Each hypothetical subject was measured on the particular dependent variable before the administration of the treatment and was again measured on the same variable after the administration of the treatment. Each pretest observation Y_{ijk} can be represented as a function of T_{ij}, a true score, plus an error term, E_{ij1}, such that the expected value of any Y_{ij1} is equal to the true score $\overline{T}_j$. Each post-test observation Y_{ijk} can be represented as a function of T_{ij}, plus the treatment effect, X_j, plus E_{ij2} such that the expected value of

Y_{ij2} is equal to $\bar{T}_j + X_j$. T_{ij} represents the true score associated with the particular observation, X_j represents the change in the true score associated with a particular treatment, and E_{ijk} represents the error associated with the particular observation of the particular subject.

The general design was to generate a random normal population that conformed to specific parameter values. These populations consisted of 6,500 observations each. Then, for each simulated experiment, there were 20 subjects or observations randomly selected from the population.

Three basic parameters were explored. Several data sets were generated such that there were differences in the amount of change caused by the treatments. Varying proportions of error variance were incorporated in the pre- and post-test scores. Furthermore, different amounts of pretest differences were represented in the data sets.

Three treatment levels were explored. In the first level, there was no difference between the means of the parent populations. In the second level, the population means differed such that the expected F ratio of the difference of means of samples taken from each of the two populations would equal 4.098, which would have an associated probability of approximately 0.05. In the third level, the population means differed from each other such that the expected F ratio of the difference of means of samples taken from each of the two populations would be 7.353, with an associated probability of approximately 0.01.

Six levels of error variance were explored. Samples were taken from populations of scores which were composed from the following levels of error variance: 0%, 10%, 25%, 35%, 50%, and 60%. The variance of the scores that was called error variance was unrelated to the variance of the scores that was regarded as true score variance. The error variance, σ_E^2, was normally distributed with the mean, $\mu = 0$. The magnitude of σ_E^2 was dependent upon the relative amount of error variance in the particular population.

In order not to confound the magnitude of error variance with the magnitude of observed score variance, the variance of observed scores σ_E^2 was maintained at a constant 1.0. The observed scores were a linear combination of true scores and error components which were respectively multiplied by their weights C_1 and C_2. If X_1 is from a population of true scores, and if X_2 is from a population of error terms, then the proportion of error variance in observed scores Y_{ij}, which are a linear combination of X_1 and X_2, can be determined by the weights of the linear combination.

If $Y_{ij} = C_1X_1 + C_2X_2$ or if $\mu_Y = C_1\mu_{X_1} + C_2\mu_{X_2}$,

then $\sigma_Y^2 = C_1^2 \cdot \sigma_{X_1}^2 + C_2^2 \cdot \sigma_{X_2}^2$.

If $\sigma_{X_1}^2 = 1$ and $\sigma_{X_2}^2 = 1$, then $\sigma_Y^2 = C_1^2 + C_2^2$.

Simply, thus if $\sigma_Y^2 = 1$, then $C_1^2 + C_2^2 = 1$.

From the above equations, if X_1 is the true score component and X_2 is the error component in the observed score Y, it can be seen that C_1^2

plus c_2^2 should equal 1.0 in all simulated conditions. The proportion of error variance, σ_E^2, was therefore equal to the square of the linear weight for the error component, c_2. Thus, the weights for the linear combination can be computed by taking the square root of the respective percentages of true score variance and error variance.

Three amounts of pretest differences of initial between-group differences were also explored. While different models or techniques may or may not be able to handle various levels of error or may introduce different kinds of distortions at different probability levels, the ultimate interest is in how well each procedure is able to evaluate change validly even though there may be initial between-group differences. In the first level, there was no difference between the means of the pretest populations being sampled. In the second level, the population means differed such that the expected F ratio of difference of means of samples taken from each of the two populations would be 4.098, with an associated probability of approximately 0.05. In the third level, the population means differed from each other such that the expected F ratio of the difference of means of samples taken from each of the two populations would be 7.353, with an associated probability of approximately 0.01.

In summary, populations were generated and subsequently sampled which met the following definitions:

Pretest control: $Y_{i11} = c_1 T_{i11} + c_2 E_{i11}$

Pretest treatment group: $Y_{i21} = c_1 T_{i21} + \pi + c_2 E_{i21}$

where T is a true score from a standard normal population, $T \sim N(1,0)$, E is the error component from a standard normal population, $E \sim N(1,0)$, α is the treatment effect, π is the initial between-group difference, C_1 is the weight for the true scores, and C_2 is the weight for the error components.

The inclusion of varying amounts of error variance is important in this kind of study. Since the social scientist operates with data that have a substantial level of error, it is of considerable importance to see how varying levels of error may influence the validity of the results of various procedures. In studies of this nature, there are various ways to interpret the meaning of error variance. In this study, the primary interpretation of error variance is that it reflects the reliability of the measuring instrument being simulated. The error levels of 0%, 10%, 25%, 35%, 50%, and 60% can be interpreted as representing respectively approximate test reliabilities of 1.00, 0.95, 0.87, 0.80, 0.70, and 0.63.

Three treatment levels, six error levels, and three pretest difference levels were utilized, thus 54 original experiments were simulated. Each time an experiment was simulated, four new populations, each of size 6,500 and each of which conformed to the above specifications, were generated. From each of these four populations,

a sample of 20 observations were randomly selected to simulate a two group, pre- and post-test experiment.

The data from each "experiment" were then analyzed using six models:

1) One-way analysis of variance of post-test scores
2) One-way analysis of variance of difference scores
3) One-way analysis of variance of residualized gain scores
4) Two-way analysis of variance
5) Repeated measures two-way analysis of variance
6) Analysis of covariance

The appropriate F ratio to evaluate the change was computed for each model for each "experiment."

The simulation of the 54 experiments was replicated a total of 50 times. Therefore, an overall total of 2,700 experiments was simulated. Across the 50 replications, the mean of the F ratios for each model used in each "experiment" was computed. These mean F ratios were used to evaluate the performance of the six models. The observed mean F ratios were statistically compared with the expected F-ratio values. Since the same "data" were analyzed with all six models, all models had a common basis of evaluation.

Results

In the cases in which the observed F ratios were anticipated to be approximately equal to the expected F ratios, such as the one-way analysis of variance of post-test scores with no pre-treatment differences, the observed mean F ratios tended to be slightly greater

than the expected value, probably because of the highly skewed nature of the F distribution. Otherwise, the results indicate that the random number generator used to create the populations worked adequately. The fluctuations from the expected values are within the range of sampling error. The mean F ratios generated by each analysis procedure in each of the simulated experiments in which there were no initial between-group differences are presented in Table 1. The mean F ratios for the simulated experiments using the second and third level of initial between-group differences are respectively presented in Tables 2 and 3.

One-way analysis of variance of post-test scores. The observed mean F ratios for the simulated experiments in which there were no initial between-group differences indicate that this procedure worked as expected. Only 1 of the 18 observed mean F ratios was significantly different from the expected value. However, when initial pre-treatment differences are present, clearly invalid and misleading F ratios are generated. All of the conditions with second and third level pretest differences resulted in observed mean F ratios for the one-way analysis of variance that were significantly different from the expected values.

Two-way analysis of variance. The fact that this model violates the assumption of independence amongst pretest/post-test observations was demonstrated by the invalid estimates of the treatment effects. The F ratios for the treatment dimension were relativley consistently lower than expected. Furthermore, the pretest differences effects

were consistently underestimated by an even larger margin. However, the value of the F ratios generated seemed to be relatively unaffected by the various levels of error variance. The treatment-pre/post interaction effects proved to be clearly invalid estimates of the change effects. For these effects, the F ratios were all significantly different from the expected values.

Two-way analysis of variance using repeated measures. This procedure does not cause the analyst to make the unwarranted assumption of complete independence amongst all the scores. Instead, there is assumed to be a dependence among the pretest and post-test scores, and therefore is represented by the inclusion of a subject dimension in the analysis. The observed F ratios were relatively unaffected by the treatment levels. For example, with no-pretest differences, at the 50% and 60% error levels, only one of the six observed F-ratio means was significantly different from the expected value. However, the amount of error variance vastly effected the validity of the observed F ratios. With no-pretest difference, the simulated experiments with 0% to 35% error produced 9 out of 12 observed mean F ratios which were significantly different from the expected value. Even at the 35% error level where the expected F ratio was 7.353, the observed mean F ratio was 11.171. At the 50% error variance level, the observed mean F ratio finally dropped to 9.389 and at the 60% error variance level, the mean F ratio was 7.487. The estimates of change effects were relatively undisturbed by initial between-group differences. Even at the third level of pretest differences, there was the

same number of observed mean F ratios which differed significantly from the expected value. However, regardless of the amount of initial between-group differences, this model was adversely affected by error levels less than 50%.

Analysis of covariance. This procedure was relatively unaffected by differences in the treatment levels. However, the F ratios generated by this procedure appeared to have been directly affected by the amount of error variance in the data. Only at the 60% level of error variance were the observed mean F ratios not significantly different from the expected values. At the 35% error variance level when the expected F ratio was 7.353, the observed mean F ratio was 13.400. At the 50% error variance level, the observed mean F ratio was 11.713, while again the expected value was 7.353. At the 60% error variance level for the third treatment level, the observed mean F ratio was 8.620. When initial between-group differences were introduced, the observed mean F ratios even further deviated from the expected values. At the third level of pretest difference, the observed mean F ratio for the 35% error level was 18.568 while the expected F ratio was 7.353. At the 50% error level for the third treatment level, the observed mean F ratio was 17.860; at the 60% error variance level, the observed mean F ratio was 17.667.

One-way analysis of variance of differences scores. This procedure produced exactly the same values for the F ratios as the two-way analysis of variance using repeated measures. Even though the mean squares produced in the two procedures had different values, the

final F ratios were exactly identical to the tenth decimal place. An example of a comparison between the two sets of results is presented in Table 4. As with the two-way analysis of variance using repeated measures, the one-way analysis of variance of difference scores proved to be unaffected by the initial between-group differences. However, the amount of error variance greatly affected the validity of the F ratios generated.

One-way analysis of residualized gain scores. The results of this procedure were very similar to those of the analysis of covariance, though generally, the F ratios computed by this procedure were slightly lower than those computed by the analysis of covariance. The value of the observed mean F ratios was greatly affected by the amount of error variance and the amount of initial between-group difference. The error variance levels of 0% to 35% resulted in 9 out of 12 observed mean F ratios being significantly different than the expected value in the no-pretest difference conditions. At the 50% and 60% error levels, 1 of the 6 observed mean F ratios was significantly different from the expected value. The introduction of between-group differences caused the observed F ratios to have even higher values, such that even at the 60% error variance level, when there were between-group differences at the third level, the observed mean F ratios were on the order of three times that of the expected F-ratio value. At the second and third level of pretest differences, 32 of 36 observed mean F ratios were significantly different from the expected values.

Discussion

These results reflect two kinds of phenomenon. The first is the ability of various statistical procedures to produce valid estimates of change effects without being disturbed by possible initial between-group differences. The second is the ability of the various procedures to validly compute change effects in the context of error variance.

The effects of initial between-group differences, or pretest effects, is the ultimate purpose of this study for it is these very effects that the models studied were designed to accommodate and overcome. The very rationale for the measurement and the evaluation of change caused by some treatment is based upon the supposition that treatment effects can be best evaluated within the context of some kind of universal baseline. It has been urged that even the most robust of between-group experimental designs ultimately contaminates the assessment of the change that occurs. Of the six procedures, only the analysis of variance of difference scores and the two-way analysis of variance using repeated measures apparently are not affected by pretest differences. These two models, which have proven to be essentially the same, apparently are able to accurately assess the treatment effects regardless of any possible biases as to initial differences between groups. The analysis of covariance model is apparently insufficient in that it results in highly inflated *F* ratios. Similarly, the analysis of variance of residualized gain scores is apparently insufficient in that it also results in highly

inflated F ratios. As expected, the one-way analysis of variance of post-test scores does not prove to be a valid way to estimate treatment effects when there are prior between-group differences. While the two-way analysis of variance model proves to be unaffected by pretest differences, it apparently produces low-estimates and, therefore, invalid estimates of the treatment effects.

Probably the most important result of this study, or insight provided by this study, is the effect due to the amount of error variance. These six models evaluate treatment effects by using a general statistical structure based upon two independent estimates of non-treatment variance (error variance) such that the two estimates differ from each other as a function of expected sampling distributions. These two independent estimates are used to form a ratio. When the observed ratio is greater than the expected ratio, then the investigator can interpret the statistic as indicating the presence of a treatment effect. In a situation where there is no error of measurement, the only sources of variance are within-group (individual-differences variance) and treatment variance. The one-way analysis of variance of post-test scores procedure is well suited to this situation. However, the elimination of pretest effects results in the removal of individual differences within each group and leaves only the between-group differences. When data are analyzed by extracting pretest values on an individual-by-individual basis, and if there is no error of measurement, then the resultant values or scores for an individual in a given group is the treatment effect

itself. In the absence of error, each individual experiences the same treatment effect. Therefore, each individual within a group has the same score. Thus, within-group variance is eliminated. When there is no within-group variance, the necessary second independent estimate of the error is unavailable. Therefore, there is no basis for a statistical evaluation.

In a condition in which there is no error, models such as analysis of covariance, one-way analysis of variance of residualized gain scores, one-way analysis of variance of difference scores, and two-way analysis of variance using repeated measures, which "control" for pretest differences and thus eliminate within-group variance, were totally unable to validly evaluate the effects of the treatments. This weakness in these models was not only apparent when the error of measurement was 10% and 25%, but also when the error was as high as 35%. The analysis of covariance procedure continued to give invalid estimates of the change effects at the 50% level of error. It should be kept in mind that a 35% error of measurement translates into a 0.806 measurement reliability.

The concern for error variance can be viewed within a wider context. The various populations, which were generated and then sampled, were defined in terms of constant treatment differences and constant pretest differences. These differences were constant in that every observation within a population differed by the same value from the observations in the other populations. The populations were further defined in terms of proportion of error (though the error had

a mean of zero) which was randomly assigned and added to each member of the populations. Thus, the populations had within-population or within-group variance that was interpreted as measurement error. Within this context, there was no treatment error in that every member of the population was affected by the treatment to the same degree. Furthermore, there were no other sources of error variance. Consideration was given also to exploring the impact of treatment error. This would reflect the more realistic situation in which a treatment affects subjects in slightly different degrees, more commonly called treatment by subject interaction. However, the inclusion of a second source of error variance would have resulted simply in a higher level of error variance, something already evaluated by the dimension of level of error of measurement. Theoretically, there would be no interactive effects between multiple sources of error since error variance is independent and has a mean of 0. A possible realistic exception is the situation in which the treatments have different levels of variance. In general, the dimension of measurement error can be interpreted within the broader context of general experimental error normally associated with educational and psychological experiments.

Summary

The results of this Monte Carlo study substantiate earlier concerns regarding the evaluation of change. The results revealed specific problems with each of the six statistical models. While the behavioral and educational researcher may be able to measure various

change phenomenon, there is now serious question as to whether or not he or she is able to statistically evaluate the change. Analysis of covariance and analysis of variance of residualized gain scores appear to be entirely inappropriate. Multiple factor analysis of variance models utilizing pretest and post-test scores appear to yield invalid F ratios. The analysis of variance of difference scores and the multiple factor analysis of variance using repeated measures are the only models which can adequately control for pre-treatment differences; however, they appear to be robust only when the error level is 50% or more. This places serious doubt regarding published findings, and theories based upon change score analysis. When an investigator is collecting data which have an error level less than 50% (which is true in most situations), then a change score analysis is entirely inadvisable until an alternative analysis model is developed.

References

Cronbach, L. J., & Furby, L. How should we measure 'change'--or should we? Psychological Bulletin, 1970, 74(1), 68-80.

DuBois, P. H. Multivariate correlational analysis. New York: Harper, 1957.

Jennings, E. Linear models underlying the analysis of covariance, residualized gain scores, and raw gain scores, presented at meetings of AERA, April 12, 1972.

Lord, F. M. The measurement of growth. Educational and Psychological Measurement, 1956, 16, 421-437.

Lord, F. M. Elementary models for measuring change. In C. W. Harris (Ed.), Problems in measuring change. Madison: University of Wisconsin Press, 1963.

McNemar, Q. On growth measurement. Educational and Psychological Measurement, 1958, 18, 47-55.

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

Table 1

Expected and Observed Mean F-ratio Values

With No Pre-Test Difference

	Expected Value	One-way ANOVA	Two-way ANOVA	Two-way Repeated Measures	Analysis of Covariance	ANOVA of Difference Scores	ANOVA of Residualized Gain Scores
Treatment Level							
				0% Error			
1	1.056	0.992	0**	0**	0**	0**	0**
2	4.098	4.297	1.636**	∞**	∞**	∞**	4,190.432**
3	7.353	8.088	3.362**	∞**	∞**	∞**	4,662.504**
				10% Error			
1	1.056	1.291	0.085**	0.821	0.879	0.821	0.868
2	4.098	3.654	1.889**	16.954**	17.171**	16.954**	16.867**
3	7.353	8.495	3.352**	32.703**	34.275**	32.703**	33.396**
				25% Error			
1	1.056	0.980	0.263**	1.020	1.073	1.020	1.078
2	4.098	4.249	1.661**	7.664**	8.662**	7.664**	8.653**
3	7.353	7.661	2.737**	15.278**	17.176**	15.278**	17.156**
				35% Error			
1	1.056	0.829	0.354**	1.025	0.825	1.025	0.815
2	4.098	4.817	1.882**	5.229**	6.355**	5.229**	6.294**
3	7.353	8.628	4.052**	11.171**	13.400**	11.171**	13.279**
				50% Error			
1	1.056	1.082	0.600**	1.146	1.164	1.146	1.153
2	4.098	5.243	2.371**	4.814	6.229**	4.814	6.293
3	7.353	9.048*	4.683**	9.389*	11.713**	9.389*	11.611**
				60% Error			
1	1.056	0.892	0.737**	1.229	0.977	1.229	0.971
2	3.098	4.733	2.212**	3.911	5.008	3.911	4.946
3	7.353	6.644	4.196**	7.487	8.620	7.487	8.629

* $p(t) < .05$

** $p(t) < .01$

where $t = \frac{\text{(observed mean F-ratio) - (expected F-ratio)}}{\text{(standard deviation of observed F-ratios)}/\sqrt{49}}$

Table 2

Expected and Observed Mean F-ratio Values With

Second Level Pre-Test Difference

	Expected Value	One-way ANOVA	Two-way ANOVA	Two-way Repeated Measures	Analysis of Covariance	ANOVA of Difference Scores	ANOVA of Residualized Gain Scores
Treatment Level							
				0% Error			
1	0.464	3.268**	0**	0**	0**	0**	0**
2	4.098	16.618**	1.649**	∞**	∞**	∞**	1,133.626**
3	7.353	19.452**	3.222**	∞**	∞**	∞**	1,433.624**
				10% Error			
1	0.464	5.735**	0.140**	1.472	1.787	1.472	1.591
2	4.098	15.502**	1.639**	16.199**	18.698**	16.199**	16.423**
3	7.353	20.424**	3.286**	31.998**	33.593**	31.998**	28.747**
				25% Error			
1	0.464	4.193**	0.232**	0.868	1.099	0.868	0.978
2	4.098	13.396**	1.835**	7.612**	11.288**	7.612**	10.220**
3	7.353	20.021**	3.965**	16.508**	22.409**	16.508**	20.169**
				35% Error			
1	0.464	4.001**	0.361**	1.119	1.753	1.119	1.636
2	4.098	16.923**	2.175**	6.098*	12.070**	6.098*	10.414**
3	7.353	19.201**	3.404**	9.459*	16.560**	9.459*	15.082**
				50% Error			
1	0.464	4.405**	0.864	1.935*	2.824**	1.935*	2.645**
2	4.098	14.421**	1.867**	3.994	9.106**	3.994	7.938**
3	7.353	17.881**	4.145**	8.737	15.259**	8.737	14.202**
				60% Error			
1	0.464	4.445**	0.626**	1.062	2.490**	1.062	2.326**
2	4.098	14.924**	2.354**	3.848	10.529**	3.848	9.460**
3	7.353	17.911**	3.571**	5.950	14.104**	5.950	12.888*

* $p(t) < .05$

** $p(t) < .01$

where $t = \frac{\text{(observed mean F-ratio) - (expected F-ratio)}}{\text{(standard deviation of observed F-ratios)}/.49}$

Table 3

Expected and Observed Mean F-ratio Values With

Third Level Pre-Test Difference

	Expected Value	One-way ANOVA	Two-way ANOVA	Two-way Repeated Measures	Analysis of Covariance	ANOVA of Difference Scores	ANOVA of Residualized Gain Scores
Treatment Level							
				0% Error			
1	1.056	9.027**	0**	0**	0**	0**	0**
2	4.098	19.371**	1.697**	∞**	∞**	∞**	664.300**
3	7.353	29.444**	3.518**	∞**	∞**	∞**	371.070**
				10% Error			
1	1.056	7.171**	0.109**	1.123	1.489	1.123	1.304
2	4.098	21.491**	1.963**	18.452**	20.833**	18.452**	16.588**
3	7.353	30.270**	3.489**	34.333**	34.995**	34.333**	25.979**
				25% Error			
1	1.056	8.809**	0.357**	1.395	2.687**	1.395	2.322**
2	4.098	21.739**	1.770**	6.959**	11.553**	6.959**	9.042**
3	7.353	26.600**	3.999**	16.276**	23.144**	16.276**	18.854**
				35% Error			
1	1.056	8.839**	0.528**	1.539	2.782**	1.539	2.339**
2	4.098	18.473**	1.596**	4.944	10.231**	4.944	8.312**
3	7.353	27.416**	3.852**	10.723**	18.568**	10.723**	15.394**
				50% Error			
1	1.056	7.944**	0.564**	1.129	3.531**	1.129	3.088**
2	4.098	20.788**	1.954**	4.202	11.835**	4.202	9.502**
3	7.353	25.636**	3.729**	7.961	17.860**	7.961	14.932**
				60% Error			
1	1.056	7.679**	0.461**	0.738*	3.311**	0.738*	2.794**
2	4.098	19.867**	2.238**	4.151	12.316**	4.151	9.863**
3	7.353	28.135**	3.169**	5.558**	17.667**	5.558**	13.675**

* $p(t) < .05$
** $p(t) < .01$

where $t = \frac{\text{(observed mean F-ratio) - (expected F-ratio)}}{\text{(standard deviation of observed F-ratios)}/\sqrt{49}}$

Table 4

Summaries of One-way Analysis of Differences Scores and Two-way Analysis of Variance With Repeated Measures

One-way Analysis of Variance of Difference Scores

	Sum of Squares	d.f.	Mean Square	F
Between	16.549	1	16.549	15.723
Within	39.996	38	1.052	
Total	56.545	39		

Two-way Analysis of Variance With Repeated Measures

	Sum of Squares	d.f.	Mean Square	F
Between				
Treatment	0.488	1	0.488	.327
Treatment x Subject	56.731	38	1.492	
Among				
Pre-Post Test	4.210	1	4.210	8.000
Treatment x Pre-Post	8.274	1	8.274	15.723
Treatment x Pre-Post x Subject	19.703	38	0.526	
Total	89.702	79		