DOCUMENT RESUME

ED 137 379 95 TM 006 185

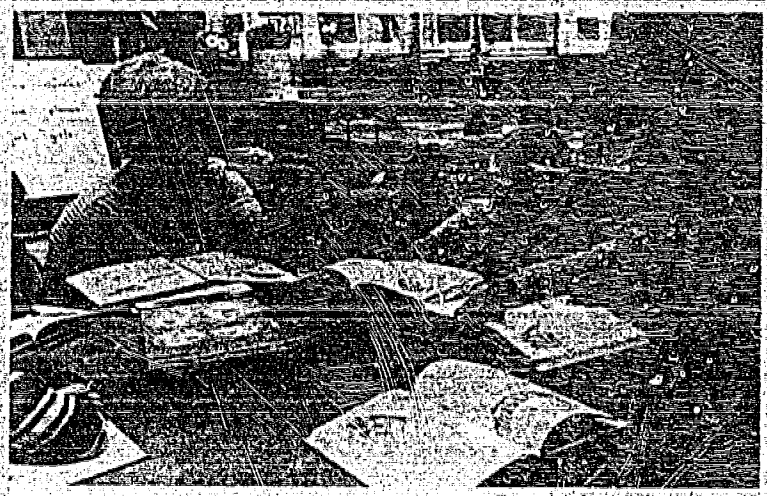
AUTHOR Marston, Paul T., Borich, Gary D.
TITLE Analysis of Covariance: Is It the Appropriate Model to Study Change?
INSTITUTION Texas Univ., Austin. Research and Development Center for Teacher Education.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE [Apr 77]
NOTE 27p.; Paper 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-$2.06 Plus Postage.
DESCRIPTORS *Achievement Gains; *Analysis of Covariance; Comparative Analysis; *Individual Differences; *Mathematical Models; Measurement Techniques; *Post Testing; Pretesting; Raw Scores; Simulation; Standard Error of Measurement; Statistical Analysis; Test Reliability; Tests of Significance; *True Scores
IDENTIFIERS Type I Error

ABSTRACT

The four main approaches to measuring treatment effects in schools; raw gain, residual gain, covariance, and true scores; were compared. A simulation study showed true score analysis produced a large number of Type-I errors. When corrected for this error, this method showed the least power of the four. This outcome was clearly the result of the computational method which adds dependent variable information into the independent variable to form the true score. Covariance analysis was recommended, with reservation, as the method of choice. (Author/MV)

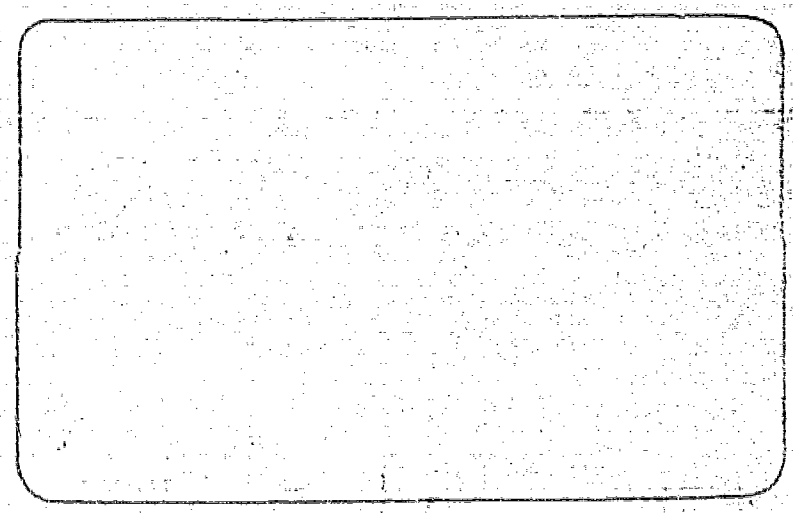
****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****



a UTR&D report

EVALUATION OF TEACHING PROJECT

M006185



The *Evaluation of Teaching Project*—one of four projects at the Research and Development Center for Teacher Education—has as its mission to develop materials and strategies for teacher training, research and evaluation. The goals of the *Evaluation of Teaching Project* are to develop (1) a conceptual framework for the evaluation of teaching, (2) a sourcebook of validated teacher evaluation and research instruments, and (3) strategies for the evaluation of teacher trainee programs. These goals are being carried out with funds from the National Institute of Education.

In the process of meeting its objectives, the *Evaluation of Teaching Project* conducts systematic research in teacher behavior for the purpose of validating instruments and identifying characteristics of effective teaching. The following report describes one facet of this research. A complete listing of studies in this report series is available by writing to the Evaluation of Teaching Project, R&D Center for Teacher Education, University of Texas at Austin, 78712.

Analysis of Covariance: Is it the Appropriate Model to Study Change?

Paul T. Marston and Gary D. Borich

Abstract

The four main approaches to measuring treatment effects in schools; raw gain, residual gain, covariance, and true scores; were compared. A simulation study showed true score analysis produced a large number of Type-I errors. When corrected for this error, the method showed the least power of the four. This outcome was clearly the result of the computational method which adds dependent variable information into the independent variable to form the true score. Covariance analysis was recommended, with reservation, as the method of choice.

Analysis of Covariance: Is It the Appropriate Model to Study Change?

Paul T. Marston and Gary D. Borich

In many testing situations, it is found that the individual differences are large relative to the size of the treatment effects being studied. What is needed is a method to compensate for the effects of these individual differences on the outcome measure so that the effect of the treatment can be accurately assessed. A number of statistical models have been designed to make this type of adjustment--usually by measuring change. Four change models are considered in this paper: (a) analysis of difference scores, (b) analysis of residual gains, (c) analysis of covariance, and (d) analysis of covariance with true score adjustment. The assumptions underlying these models are examined and then the results of a Monte-Carlo simulation study comparing the four methods is reported.

All of these methods start with the assumption that an individual's posttest score, y, can be thought of as a linear combination of a number of factors including the initial level of performance. In the case of the single treatment being considered here, the even stronger assumption is made that the only important factors are the effect of the treatment, a, and the individual's pretest performance, x. The critical differences between these models involve the assumptions made about the relationship between the pretest score and posttest score. All four models assumed that the relationship is linear and is the same for all individuals.

The method which makes the strictest assumptions is the analysis of difference scores. In it the posttest score is thought to consist of a treatment effect plus the individual's pretest score. In other words, everyone in a specific treatment group is expected to change by the same amount. The

usual approach in making an analysis of difference scores is to form a gain score by subtracting each individual's pretest from their posttest score. Then a fixed effect analysis of variance or a t-test is made on these gain scores. The two trial, mixed model analysis of variance is also used for pretest-posttest designs, but as Huck and McLean (1975) have pointed out, this analysis of variance is formally identical to the analysis of difference scores.

The assumption of a constant change for every individual in a group does not appear to be a reasonable one for many situations. When posttests scores are plotted as a function of their respective pretest value, it is usual to observe regression toward the mean. That is, an individual with a high initial score is more likely to score lower on the posttest and conversely, an individual with a low pretest scores is likely to raise their score. This means that the amount of change is a function of the initial level of performance when everything else is held constant. If the regression of the posttest scores on the pretest is reasonably linear then an estimate of the posttest score can be made by using a simple correlation model. Gain scores are then formed by subtracting an individual's estimated posttest score (formed using the regression equation with pretest) from their actual posttest scores. Treatment effects would then show up as different mean gains for the various treatment conditions. Such a procedure is called residual gain analysis. As a technique, it gets around the assumption of equal gains for all individuals regardless of initial score used in the analysis of difference scores, while still retaining the type of computation and interpretation found in the latter type of analysis. One should bear in mind that the average gain across all groups will be zero in the residualized gain method.

In some ways residualized gain is very similar to analysis of covariance which is examined next. Both appear to be using the following underlying linear model for estimating the posttest scores;

$$y = b_0 + b_1 a + b_2 x + e \quad . \qquad (1)$$

Where a and x are the treatment and pretest effects, e is a random error, and the b's are weighting coefficients. What distinguishes the types of analyses is how the estimates of the weighting coeffecients are obtained. The analysis of covariance makes the fewest assumptions by allowing all three b's to be fitted from the data. Werts and Linn (1970) have shown that in the residualized gain analysis the value of b_2 is assumed to be the same as what would have been obtained if the a term was not included in the model. They also show that in the difference score analysis the assumption is that the value of b_2 is equal to 1.0. So it is clear that the difference score analysis also assumes Model (1). One can thus think of the three analyses as putting progressively less restrictive assumptions on the same theoretical model. For a given data set, analysis of covariance can never give a worse representation of the relationships than the other two and it may often be better. This is because the least-squares solution for the b's involves the interrelationship of all the variables. For example, the value of b_2 in the two group cases depend on the correlations r_{xy}, r_{ay}, and r_{ax}. In the residualized gain analysis the value of b_2 can only be a function of correlation r_{xy}. Werts and Linn show that the two methods will give equivalent results only if there is no correlation between the covariate and the treatment variable. Given the additional work of forming the residualized gain scores, it is not clear why there would ever be a preference for what is only an approximation to analysis of covariance.

As the assumptions on the theoretical model are progressively relaxed in the three methods discussed so far, a better estimate of b_2 is obtained and consequently a better estimate of the treatment effect, b_1, is also obtained. If this line of reasoning is continued, it appears that when the analysis of covariance does not give a good estimate of b_2 then an improved model should be

sought. This can happen when errors of measurement occur in obtaining the x values. Such errors will lower the obtained correlation between the pretest and posttest relative to that which would have been obtained if accurate measurements had been used. When such accurate measurements in principal cannot be obtained then the hypothetical accurate values the x's represent are called "true scores." The appropriate analysis using these scores is logically enough called true score analysis of covariance or true score analysis for short. Students of measurement theory have argued that when the covariate contains error its correlations with the other variables should be corrected for the unreliability before the analysis is done (Cronbach & Furby, 1970). The "true" pretest-posttest relationship can be obtained and therefore a better estimate of the treatment effect is also obtained. Such corrections are based on some reliability measure for the pretest such as a test-retest intraclass correlation.

Up to this point there appears to be little disagreement in the literature as to the merits of the first three models discussed. It is recognized that the covariate may not relate to the dependent variable in a linear fashion or even if it does this relationship might be a function of the treatment. Both of these assumptions can and should be tested prior to making tests using the analysis of covariance model (Draper & Smith, 1966). The assumption that a correction should be made for an error of measurement in the covariate is far from universally accepted. For example, writers of one textbook state that it does not make any difference whether the independent variables are measured with or without error (Draper & Smith) while another simply says it limits one to making statistical inferences about the obtained scores (Graybill, 1961). Lumsden (1976) has taken the position for ignoring the reliability question altogether. He states that the true score itself can be considered as an unreliable measure of some actual characteristic of the individual. For

example, one ought to be interested in the relationship between the obtained scores on a math test and mathematical ability, not in the relationship between obtained math scores and true math scores. In other words, why stop at correction for test reliability? Why not correct for the error of measurement between the test and the individual characteristic? The latter obviously cannot be done so why bother with the former. This substantive criticism should be borne in mind when considering the other evidence about the true score adjustment.

The treatment of corrections for an analysis of true scores is based on a least-squares solution from an adjusted intercorrelation matrix (Cohen & Cohen, 1975; Cronbach & Furby, 1970; Werts & Linn, 1970) so it is not always clear what the estimated true scores would be. One can find out, however, by using the adjusted correlation matrix to solve the equation for the true score, $\underline{x}_t$;

$$x_t = c_0 + c_1 x + c_2 y \quad . \tag{2}$$

This solution gives

$$x_t = \left(\frac{\sqrt{1 - \frac{r_{xy}^2}{R_{xx}}}}{\sqrt{1 - r_{xy}^2}} \right) x + \left(\frac{r_{xy}}{\sqrt{R_{xx}}} - r_{xy} \frac{\sqrt{1 - \frac{r_{xy}^2}{R_{xx}}}}{\sqrt{1 - r_{xy}^2}} \right) y \quad . \tag{3}$$

If these new $\underline{x}_t$ values are substituted for the original $\underline{x}$'s in the data, one finds the new correlation is

$$r_t = \frac{r_{xy}}{\sqrt{R_{xx}}} \tag{4}$$

which is the correction for unreliability. When a treatment effect is added to equation (2) and solved it is found that $\underline{x}_t$ is a function of both $\underline{y}$ and $\underline{a}$.

That result is somewhat complex and will not be included here. The key thing to note is that in obtaining the true score values, one winds up using the posttest to predict itself--a procedure with a certain amount of circularity. In fact, if r^2_{xy} is equal or greater than the reliability, R_{xx}, then the true scores will perfectly predict the posttest values. If this happens, then no matter how large the treatment effect is, it must always be estimated as zero in the true score analysis. This problem could be circumvented by switching to a standard analysis of covariance whenever the magnitude of r^2_{xy} is close to the reliability correction. Now one only has the problem of deciding how close is really close.

The Simulation Study

A Monte-Carlo type simulation study was designed to shed some light on the relative merit of these four methods of measuring change. The baisc population model was a two group experiment with a linear relationship between the pretest and posttest. The linearity between pretest and posttest scores was held constant throughout the sample sets at a correlation of .6. This value permitted the reliability to be varied over a wide range while still giving a fair amount of power for the covariate. Analysis of covariance, true score analysis, analysis of residualized gain, and analysis of difference scores were calculated for each sample and the number of significant F-tests at four standard levels were tabulated.

To generate the pairs of pretest and posttest observations, a set of three random normal deviates was required. The value of the pretest was set to the first random normal deviate, e_{1ij}. The posttest score was then calculated using the relationship

$$y_{ij} = m + r_{xy}x_{ij} + g_i + \sqrt{1 - r^2_{xy}}\ e_{2ij} \quad . \qquad (5)$$

The parameters $\underline{m}$ and $\underline{g}_i$ represent the grand mean and the deviation score for group, respectively. To simplify calculations, $\underline{m}$ was set equal to zero. The obtained pretest with error, $\underline{x}'$, was obtained using the relationship

$$x'_{ij} = \sqrt{R_{xx}}\, x_{ij} + \sqrt{1 - R_{xx}}\, e_{3ij} \quad . \tag{6}$$

Every sample had two groups of twenty-five observations and there were 1000 samples in each set. Reliability values, R_{xx}, were set a .8, .6, and .5 while differences between the groups on the dependent variable, $\underline{y}$, were set at .6, .2, .1, and 0.0. The latter, of course, represents a test of the null hypothesis. Additionally, a few simulations were performed with differences introduced into the means of both the pretest and posttest groups such that the predicted $\underline{y}$ values lay on the same regression line. Because the unreliability was added in a second step, it was possible to obtain an actual score analysis of covariance by using the $\underline{x}$ values instead of the $\underline{x}'$ values. In some of the later simulations this was done. The four levels for alpha used were: .01, .05, .10, and .25.

After reviewing the first few simulations the program was modified to provide descriptive statistics for each of the methods and to provide an analysis of covariance on the reliable $\underline{x}$ scores.

The true score correction was one suggested by Cohen and Cohen (1975). It requires a least-squares procedure that uses the correlation matrix to obtain a solution. The raw score correlation matrix is altered by dividing all correlations involving the unreliable variable by the square root of R_{xx} and multiplying the standard deviation of the unreliable variable by the same value. The least-squares solution is then found for this new matrix. Inspection of the $\underline{b}_2$ weights for the true score analysis indicated the correction may have been too large relatively to the amount of unreliability introduced in the model.

Some additional sample sets were then produced with the size of the unreliability correction being set to half the amount of unreliability introduced. For example, if the reliability of x in equation (6) was set at .8, then the appropriate correlations in the raw score matrix were divided by the square root of .9.

Results

The most interesting aspect of the simulations was the distribution of Type-I errors. Table 1 presents the average number of significant results for the various types of analyses. Three parameters are used to make a chi-squared goodness of fit test for an F distribution (i.e., number of samples, df_1, and df_2) so a chi-squared test with one degree of freedom could be conducted on the number of significants F's. All but one of the true score distributions could be rejected as a bad fit at $p < .01$ while none of the covariance, residualized gain, or difference score distributions could be rejected at this level. This test clearly indicated the true F-tests were not following the expected distribution of Type-I errors. An estimate of the relationship of true score alpha level to that used in the tests was found by fitting a second degree polynomial with a zero intercept to the number of significant true score F-tests as a function of the expected number of significant F-tests. Figure 1 shows the plot for a reliability of .6. With the exception of the $p < .01$, this curve appears to be a reasonable fit to the values. The half size reliability corrections were a good match to the corresponding plot for a reliability of .8. It appears that the primary effect of the true score correction was to increase the effective alpha level for the F-tests. Table 2 gives the estimated alpha level values for true score correction based on the polynomial equations.

When the groups had mean differences on the posttest the true score analysis did produce more significant F-tests than analysis of covariance. The difference in alpha levels of the two methods, however, made direct

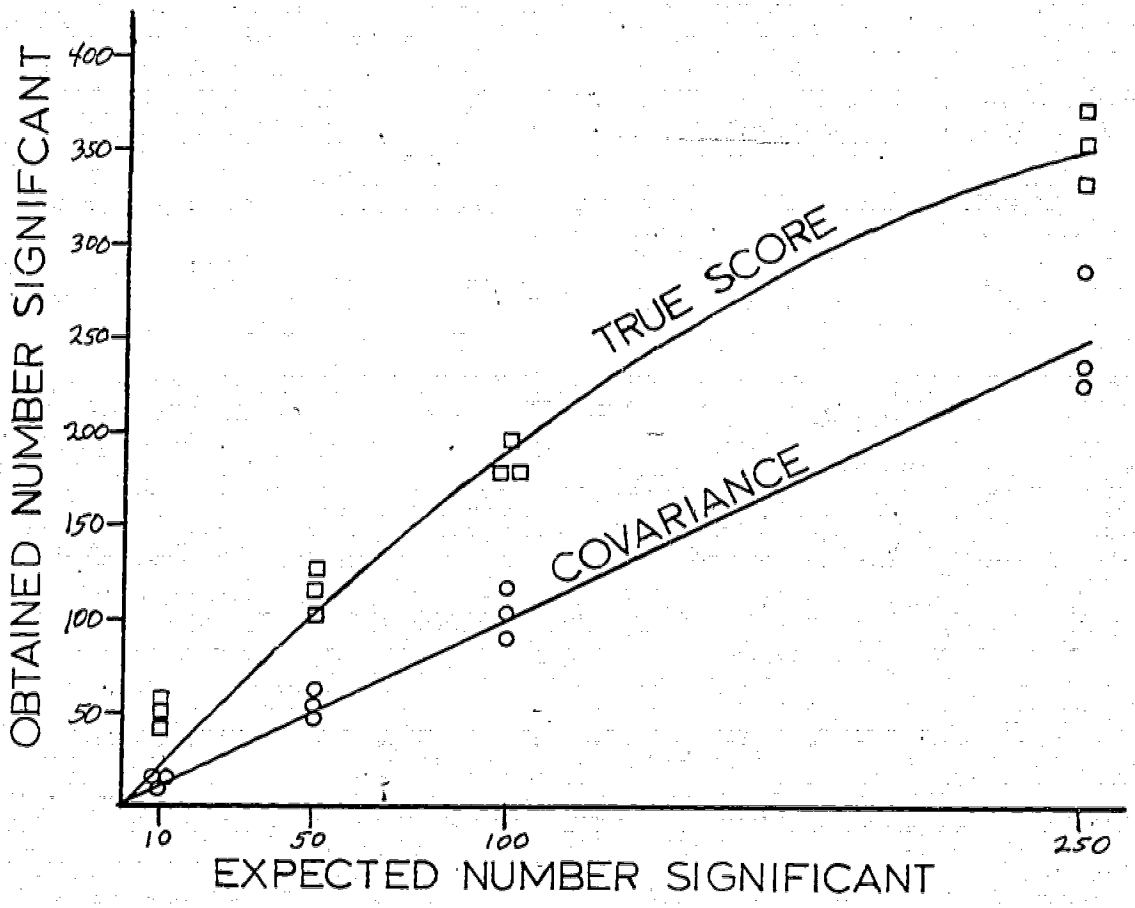
Table 1

Number of significant F-test obtained when the pretest and posttest means have no differences.

Expected Number Significant	Error in X	Correction R_{xx}	Number of Sample Sets	Covariance	True Score	Residual Gain	Difference Scores
250	.8	.9	1	239	254	238	238
	.8	.8	3	259	291	258	259
	.6	.8	1	233	268	234	258
	.6	.6	3	250	320	250	254
	.5	.5	2	240	382	239	245
100	.8	.9	1	91	100	89	89
	.8	.8	3	105	135	105	107
	.6	.8	1	88	109	90	91
	.6	.6	3	102	182	101	92
	.5	.5	2	96	226	97	101
50	.8	.9	1	44	52	44	43
	.8	.8	3	52	76	50	53
	.6	.8	1	47	61	47	48
	.6	.6	3	53	112	55	48
	.5	.5	2	45	152	46	48
10	.8	.9	1	5	8	5	10
	.8	.8	3	9	20	9	12
	.6	.8	1	12	18	13	8
	.6	.6	3	11	46	10	12
	.5	.5	2	10	65	11	11

Figure 1

Distirbution of significant F-tests for True Score Analysis and Analysis of Covariance when the null hypothesis is true.



comparison impossible. By plotting the number of significant F- tests as a function of the estimated alpha level, the power of the true score method could be compared with analysis of covariance. Figure 2 shows the relative power for the two types of analyses when the reliability was .6. The analysis of covariance appears to be slightly more powerful because at nearly every level of alpha it produces more significant results. This would indicate that the power of a true score analysis could be obtained more directly by increasing the alpha level in analysis of covariance.

It is possible that true score analysis might do a better job of recovering the population parameters of the model. Table 3 gives the mean and standard deviation for each b weight calculated in the two types of analyses. As expected, introducing unreliability into the pretest scores caused the covariance analysis to get a lower value for b_2. The true score analysis got a larger value for b_2 when the full correction was introduced and came very close to the actual value when the half correction was used. Even so, examination of the means for b_1 and b_0 shows that the two methods produced almost identical estimates and that these were also almost identical to the mean weights before the unreliability was introduced. There was a difference in the standard deviations of the weights for the two methods. In almost every case, the true score analysis produced a greater variation in the estimation of each of the three weights than did the analysis of covariance. In terms of estimating the critical parameter b_1, the true score analysis appears to do no better than analysis of covariance and in fact may be worse judging by the standard deviations of the weights.

The one place where the true score analysis did appear to have an edge was that case when there was a mean difference in both the pretest and posttests for the groups. Table 4 shows that, when the half unreliability correction was used, the true score sample sets came closer to following the

Table 2

Estimated Type-I error rates for
true score analysis of covariance

Alpha Used	(True score correction)/(unreliability introduced) .9/.8	.8/.8	.8/.6	.6/.6	.5/.5
.01	.014	.015	.015	.022	.030
.05	.058	.073	.073	.104	.136
.10	.110	.138	.138	.191	.242
.25	.275	.290	.290	.352	.380

Figure 2

Distribution of significant F-tests for True Score Analysis (TS) and Analysis of Covariance (CV) when the null hypothesis if false.

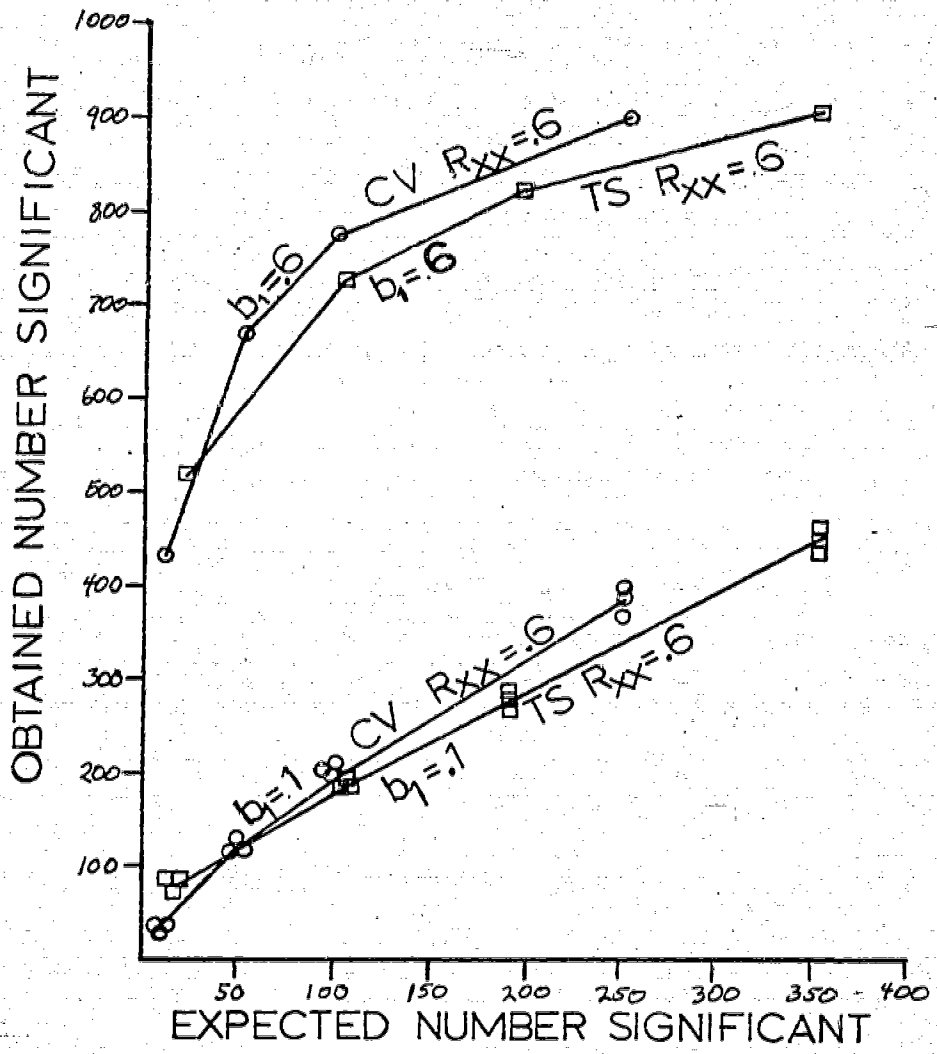


Table 3(a)

Mean values of sample regression weights

Reliability	Correction	b_0			b_1			b_2		
		AS	CV	TS	AS	CV	TS	AS	CV	TS
					$Y = 0 + .3a + .6x$					
.8	.9	-.001	.001	.001	.302	.298	.298	.601	.534	.595
.8	.8	-.005	-.005	-.006	.294	.294	.294	.604	.541	.680
.6	.8	-.001	-.001	-.001	.310	.311	.310	.599	.465	.584
.6	.6	.000	-.002	.001	.310	.313	.313	.595	.456	.772
					$Y = 0 + .1a + .6x$					
.8	.9	.000	.001	.001	.097	.096	.095	.600	.538	.599
.8	.8	.003	.001	.001	.098	.099	.100	.600	.537	.675
.6	.8	.009	.007	.007	.102	.104	.104	.595	.462	.581
.6	.6	.007	.006	.007	.099	.098	.099	.593	.454	.768
					$Y = 0 + .05a + .6x$					
.8	.8	.007	.007	.007	.042	.039	.039	.604	.541	.681
.6	.6	.000	.003	.003	.055	.050	.052	.603	.468	.792

Ledgen: AS Actual Score Analysis of Covariance

CV Analysis of Covariance

TS True Score Analysis of Covariance

Table 3(b)

Standard deviation values
for sample regression weights

Reliability	Correction	b_0			b_1			b_2		
		AS	CV	TS	AS	CV	TS	AS	CV	TS
					$Y = 0 + .3a + .6x$					
.8	.9	.112	.119	.119	.118	.122	.123	.120	.130	.145
.8	.8	.111	.118	.121	.115	.122	.125	.114	.123	.154
.6	.8	.111	.130	.133	.118	.134	.135	.117	.131	.164
.6	.6	.117	.132	.145	.116	.129	.141	.119	.133	.226
					$Y = 0 + .1a + .6x$					
.8	.9	.114	.120	.121	.111	.119	.120	.114	.120	.133
.8	.8	.112	.118	.120	.116	.122	.126	.116	.125	.156
.6	.8	.111	.124	.126	.114	.127	.129	.119	.131	.165
.6	.6	.110	.123	.129	.112	.125	.136	.116	.129	.219
					$Y = 0 + .05a + .6x$					
.8	.8	.115	.120	.122	.112	.118	.121	.117	.125	.158
.6	.6	.113	.124	.136	.115	.127	.139	.118	.129	.220

Ledgen: AS Actual Score Analysis of Covariance

CV Analysis of Covariance

TS True Score Analysis of Covariance

null distribution. Too many significant results were still obtained and the number of these appeared very sensitive to the size of the reliability correction. The analysis of covariance of the actual scores under these conditions did result in a null distribution of significant F-tests. This set of samples substantiates the warning that covariance may not be appropriate for data sets where there are large group differences in the means of the covariate. As Lord (1963) points out, one should make every effort to keep this from happening by techniques such as random assignment of subjects to groups.

While a simulation can never directly answer theoretical questions about statistical models, it does give important clues as to what important factors might be. For instance, why should the true score analysis produce so many Type-I errors? This probably happens because the addition of dependent variable information into the true score predictors increases the R^2 for the model as a whole. While most of this increase goes into the b_2 component in some samples it also gets into the other weights resulting in an excessive number of significant findings. There is, of course, no way of knowing when this will happen. To make matters worse, it is often very difficult to determine just what the value of R_{xx} should be. Some of our own data indicates that it can vary by large amounts from group to group, particularly for groups occurring naturally. It is not clear at all how one incorporates multiple values of R_{xx} into the reliability correction model.

Very little has been said about the other two methods of analyzing change, the analysis of difference scores and the residualized gain analysis. In general, they did just as one would expect from Werts and Linn's paper. The distribution of Type-I errors paralleled that of analysis of covariance when the null hypothesis was true and both methods showed less power when it was not true. The sampling method used apparently produced very little correlation between the covariate and the group factor because the residualized gain

Table 4

Number of significant F-tests when the pretest means are different and the posttest means lie on the same regression line (R_{xx} = .6)

Y mean difference	Correction	Test	Expected number significant 10	50	100	250
.6	.8	TS	31	104	175	346
		CV	48	143	221	410
.6	.8	TS	118	203	266	428
		CV	44	127	211	409
.2	.6	TS	57	132	207	372
		CV	14	57	104	266

F-distributions were close to covariance in almost every case.

Clearly, the analysis of covariance is the method of choice to control for individual differences on a posttest measure. To some extent it will also control for group mean differences on the covariate but there are problems with this. Again, the researcher is cautioned to make tests for homogeneity and linearity of regression a standard procedure. The homogeneity of regression slopes test is especially important when there are mean pretest differences in the groups.

References

Cohen, J. & Cohen P. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Erlbaum, Hillsdale, New Jersey: 1975.

Cronbach, L. J. & Furby, L. How should we measure "change"--or should we? Psychological Bulletin, 1970, 75, 68-80.

Draper, N. R. & Smith, S. Applied Regression Analysis. Wiley, New York: 1966.

Graybill, F. A. An Introduction to Linear Statistical Models, Vol. 1. McGraw-Hill, New York: 1961.

Huck, S. W. & McLean, R. A. Using a repeated measures ANOVA to analyze data from a pretest-posttest design: a potentially confusing task. Psychological Bulletin, 1975, 82, 511-518.

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

Lumsden, J. Test theory. Annual Review of Psychology, 1976, 27, 251-280.

Werts, C. E. & Linn, R. L. A general linear model for studying growth. Psychological Bulletin, 1970, 73, 17-22.