A comparison of various regression-correlation methods for evaluating nonexperimental research

David Marcus Rindskopf
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

Part of the Psychology Commons

Recommended Citation
Retrospective Theses and Dissertations. 5696.
https://lib.dr.iastate.edu/rtd/5696

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or “target” for pages apparently lacking from the document photographed is “Missing Page(s)”. If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in “sectioning” the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from “photographs” if essential to the understanding of the dissertation. Silver prints of “photographs” may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

Xerox University Microfilms
300 North Zeeb Road
Ann Arbor, Michigan 48106
RINDSKOPF, David Marcus, 1948-
A COMPARISON OF VARIOUS REGRESSION-CORRELATION METHODS FOR EVALUATING NONEXPERIMENTAL RESEARCH.

Iowa State University, Ph.D., 1976
Psychology, general

Xerox University Microfilms, Ann Arbor, Michigan 48106
A comparison of various regression-correlation methods for evaluating nonexperimental research

by

David Marcus Rindskopf

A Dissertation Submitted to the Graduate Faculty in Partial Fulfillment of The Requirements for the Degree of DOCTOR OF PHILOSOPHY

Major: Psychology

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

Iowa State University

Ames, Iowa

1976
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Regression artifacts</td>
<td>4</td>
</tr>
<tr>
<td>Correcting for unreliability</td>
<td>8</td>
</tr>
<tr>
<td>An alternative to the pretest</td>
<td>18</td>
</tr>
<tr>
<td>Factor analysis methods</td>
<td>20</td>
</tr>
<tr>
<td>Combining factor analysis and regression</td>
<td>20</td>
</tr>
<tr>
<td>Fitting factor models</td>
<td>27</td>
</tr>
<tr>
<td>Summary</td>
<td>31</td>
</tr>
<tr>
<td>SIMULATED DATA</td>
<td>32</td>
</tr>
<tr>
<td>Method</td>
<td>32</td>
</tr>
<tr>
<td>Results</td>
<td>33</td>
</tr>
<tr>
<td>REAL DATA</td>
<td>37</td>
</tr>
<tr>
<td>Method</td>
<td>37</td>
</tr>
<tr>
<td>Head Start</td>
<td>38</td>
</tr>
<tr>
<td>Title 1</td>
<td>42</td>
</tr>
<tr>
<td>Results</td>
<td>46</td>
</tr>
<tr>
<td>Title 1 Metropolitan test</td>
<td>46</td>
</tr>
<tr>
<td>Title 1 Stanford test</td>
<td>48</td>
</tr>
<tr>
<td>Head Start Blacks</td>
<td>50</td>
</tr>
<tr>
<td>Head Start whites</td>
<td>51</td>
</tr>
<tr>
<td>Double pretests</td>
<td>51</td>
</tr>
<tr>
<td>Fitting factor models</td>
<td>55</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>57</td>
</tr>
</tbody>
</table>
Introduction

There has been much debate about proper methods for analyzing the results of nonexperimental studies. In nonexperimental studies random assignment of subjects to control and treated groups is absent, thus eliminating the best method of attempting to achieve equivalence prior to the start of treatment. The absence of randomization most often occurs in the most important studies -- those "real world" studies which may be used to make broad policy decisions, such as which method of therapy, education, or social welfare system is the best to implement. Since many researchers consider it either impossible or impractical to conduct true experiments in these nonlaboratory settings (but see Boruch, 1976, for a dissenting opinion), it is imperative to evaluate various methods which attempt to minimize the possible biases arising from pre-existing differences between groups in these studies. In connection with this problem, the purposes of this paper are to:

(1) briefly outline the problems which have been noted in evaluating nonexperimental studies and the methods presently used for solving these problems, and

(2) describe some methods of correlation and regression analysis which try to solve these problems, and apply these methods to both real and simulated data sets in order to evaluate them.
When faced with a body of data from a nonexperimental study, a data analyst has two choices. One choice is to say that the data are worthless since assignment to groups was not random, and refuse to do any analysis of it. Keppel (1973), for example, in discussing loss of subjects from a study asks, "Has the loss of subjects, for whatever reason, resulted in a loss of randomness? If it has, either we must find a way to restore it or simply junk it. No form of statistical juggling will rectify this situation." The other possible choice is to go ahead and analyze the data, keeping in mind that since assumptions of the various statistical tests may have been violated, the conclusions reached must be more tentative than when the same methods are applied to data from a true experiment. The results, nonetheless, may add some valid information to help determine whether the treatment had an effect.

If a problem is important enough so that even imperfect data should be analyzed, then it might also be advantageous to perform many different analyses of the same data set. Since most alternative ways of analyzing data involve making different sets of assumptions, then if the alternative methods give the same answer we can have greater confidence in our results. If the different methods give different answers, we will have to be more careful, but may be able to make plausible inferences about which assumptions were
likely to have been violated, and thus which analyses would be most accurate. Furthermore, if the various methods we use have overlapping sets of assumptions, we may even be able to tell which assumptions we are most likely to have violated by looking at the patterns of results. For example, suppose that Method A gives a large number of Type I errors when there is a violation of assumptions X and Y, while Method B gives a large number of Type I errors when there is a violation of assumptions X and Z. If, on a particular data set, the null hypothesis is rejected for Method A but not for Method B, then we have evidence that assumption Y was violated, but assumption Z was not. Therefore Method B is most probably giving the correct result in this application.

Campbell and Boruch (in press) have discussed six ways in which nonrandom assignment can lead to improper conclusions: regression artifacts, floor and ceiling effects, differential growth rates, changes in reliability of tests, between group differences in reliability, and grouping feedback effects. The regression artifact has been the most extensively discussed of all these, and is probably the most pervasive in terms of affecting the greatest number of studies. Since the regression artifact is the principal problem which the correlational and regression methods are sensitive to, the nature of this artifact and methods used
to adjust for it will be explained in detail.

**Regression artifacts**

Although the regression artifact was documented many years ago (Rulon, 1941, and Thorndike, 1942, for example) it was brought most forcefully to the attention of researchers by Campbell (Campbell & Stanley, 1966; Campbell & Erlebacher, 1970, Campbell & Boruch, in press). Campbell's approach to describing the problem is basically the one I will use, along with some examples from Lord (1960).

To begin with, consider a single group of people who are measured at two different times on the same characteristic. Either due to changes in the people themselves or errors of measurement, the correlation between the two measures will probably be less than 1.0, and for most variables used in psychological studies it will be well below 1.0. An examination of a scatterplot of the measurements will reveal that low scorers on the first measurement will tend to get better scores on the second measurement, while higher scorers will tend to decline (assuming equal means and variances for both occasions of testing). If the scores are bivariate normal, with equal means and variances, the equation for predicting score 2 from score 1 is \( \hat{X}_2 = \bar{X} + r(x_1 - \bar{X}) \), where \( r \) is the correlation between score 1 and score 2. Note that if \( X_1 \) is less than the mean, the predicted score 2 (which is
equivalent to the average score 2 for all those who got $X_1$ on the first test) will be greater than $X_1$ but less than $\bar{X}$. Thus, the second score will be expected to be closer to the mean (regressed toward the mean). Although the implicit assumption made in the previous line of reasoning is that $r$ is positive, the conclusion still holds if $r$ is negative.

This example of regression toward the mean can be used to illustrate the fallacy of using gain scores to evaluate treatments given to selected groups. The mean of any group selected because of its extreme score on a test will tend to regress toward the mean on a second test, even if no treatment is given, or if an ineffective treatment is given. Thus, a group of "underachievers," selected because they scored lower than average on a test, would be expected to do better on a second test merely because of the imperfect relationship between scores on the two tests. Thus, studies showing that a particular program increased test scores of originally low scoring students could be totally discounted as being a predictable artifact.

Now consider a typical nonexperimental design comparing a control group with a treated group. Suppose that the treated group was given the treatment because they showed the greatest need, perhaps by scoring low on some pretest. What would the results be if the treatment were totally ineffective? If it is assumed that the control and treated
groups are samples taken from two different subpopulations, each with different characteristics, then each subject would regress towards his group's mean. If group means were compared, there would be no difference in gain scores from time 1 to time 2 for either group, thus leading to the correct conclusion that the 'treatment' was ineffective. But if various methods are used to try to 'adjust' for the pre-existing differences between groups, problems can arise.

For example, if matching is attempted, the only matches will come from the highest scoring members of the treated group who would be matched with the lowest scoring members of the untreated group. But if each person's score regresses towards the subgroup mean, the control subjects will tend to have higher scores on the second test (they will regress upwards), while the treated subjects will regress downwards toward their subpopulation mean. This will make the treatment seem harmful, when in fact it is ineffective. These results would occur even if there were no treatment.

One of the most widely used statistical techniques in educational research is analysis of covariance (ANCOVA). By using this technique, researchers hope to correct for initial differences between subjects. In educational situations the application of the technique usually leads to biased results because the covariate is measured with error
(i.e., the covariate is not perfectly reliable), and the control and experimental groups are usually not equivalent.

Lord (1960) gave an intuitive demonstration of how error in the covariate combined with using nonequivalent groups can lead to inappropriate conclusions. Figure 1 (adapted from Lord's article) shows two groups of subjects plotted by score on a pretest and a posttest. Assume that group A is the control group and group B is the experimental group. It appears that the difference between the groups must be at least partially due to harmful treatment effects, because the regression line for group A is parallel to and has a higher intercept than that of group B. Usually, the differences in intercepts would be used as a measure of treatment effect.

In Figure 2 (also adapted from Lord's article) this 'treatment effect' is shown to be an artifact. Assume that the straight line which intersects the y axis at point P denotes the true relationship between pretest and posttest. In this case, although groups A and B would differ on both pretest and posttest scores, the regression line describing the relationship would be the same for both groups, thus indicating no treatment effect. The dashed horizontal lines indicate the effect of error in the covariate, which is to spread people out horizontally from their true score on the covariate x. The results is to produce scatterplots of
groups A and B identical to Figure 1. Lines aa and bb show the regression of y on x for these groups now that error has been introduced. The error produces a difference between the intercepts of the regression lines of the two groups, thus making it appear that the 'treatment' had an effect.

This problem of error in covariates is a real one in educational research; in fact it is almost inescapable. In most studies some measure of aptitude or achievement (usually a standardized test or socioeconomic status indicator) is used as a covariate. If internal consistency is used as an estimate of reliability, the upper limit is usually about .90, and if test-retest reliability is used the range may be anywhere from .50 to .85 depending on the particular test used, the sample of subjects, and the amount of time between testings. Although there is some disagreement about which type of reliability estimate is appropriate in which situations, it is clear that conclusions can be drastically altered when corrections are made, as will be seen later.

**Correcting for unreliability**

There are two basic ways in which the corrections can be made: A nearly unbiased estimate of true scores can be computed and then inserted into an analysis of covariance, or the corrections can be made to the sums of squares directly, without bothering to compute estimated true scores.
for each subject. Both the Porter (1967) and the DeGracie-Fuller (1972) methods described below use the latter approach of correction of sums of squares, but they use slightly different estimators and thus get slightly different results. When reliability estimates are low, the methods give different answers; and for reasons to be discussed later, the Porter formula is assumed to be more likely to be valid. Estimates using the DeGracie-Fuller method will be listed for some analyses so they may be compared with the estimates from the Porter method.

In the following section I will demonstrate the use of the Porter and DeGracie-Fuller methods with one set of data from one of the studies analyzed in this report.

These data are from the nationwide evaluation of Title I compensatory education programs. The scores were grade equivalent scores on the Metropolitan test, a standardized achievement test. The data necessary for the traditional ANCOVA and both the Porter and DeGracie-Fuller corrections are contained in the following data:
The usual corrected within sum of squares for an ANCOVA can be computed from the within-groups total line as follows:

\[ SS_w = \Sigma y^2 - (\Sigma xy)^2/\Sigma x^2 = 79193.27 - (31367.12)^2/30982.87 = 47437.14; \ \text{df}_w = N-3 = 955; \ MS_w = SS_w/\text{df}_w = 49.67. \]

The sum of squares for deviations about the regression line can be computed using the SS and information from the Total line:

\[ SS_B = SS_T - SS_w = 81409.50 - 47437.14 = 34162.36 \]
\[ SS_B = SS_T - SS_w = 81409.50 - 47437.14 = 34162.36 \]

In a conventional ANCOVA, the F ratio for testing the null hypothesis would be

\[ F = MS_B/MS_w = 679.11/49.67 = 13.67 (p < .001). \]

Thus a conventional ANCOVA would indicate that there were treatment effects. An examination of the regression lines for the control and experimental treatment groups would show that the experimental group seems to be inferior to the
control group.

If we now add the assumption that pretest scores may be measured with error, so that the reliability (R) equals true score variance divided by observed score variance, Porter's technique gives the following adjustment for the total sum of squares

\[
SS^*_T = SS^*_Y - \{(R) (SS_{wXY}) + (SS_{BXY})\}^2 / (R^2 SS_{wX} + SS_{Bx}),
\]

where

\(SS^*_T = \) corrected total sum of squares

\(SS^*_Y = \) sum of squares for \( y \) from total line = 81409.50

\(SS_{wXY} = \) within-groups sum of cross-products = 31367.12

\(SS_{BXY} = \) between-groups sum of cross-products = 968.63

\(SS_{wX} = \) within-groups sum of squares for \( x \) = 30982.87

\(SS_{Bx} = \) between-groups sum of squares for \( x \) = 422.91

\(R = \) reliability estimate (test-retest correlation) = .60

The within-groups sum of squares stays the same as in the conventional ANCOVA, so the adjusted between groups sum of squares is given by

\[
SS^*_B = SS^*_T - SS^*_W
\]

Thus, for this example, where the test-retest correlation of .60 was used as the reliability estimate, we have

\[
SS^*_T = 81409.50 - \frac{((.60) (31367.12) + 968.63)^2}{(.60)^2 (62286.21) + 4171.44} = 47583.01
\]

\[
SS^*_B = 47583.01 - 47437.14 = 145.87
\]

The Porter adjusted F ratio is
Note that in making the correction for unreliability, the $F$ statistic computed is not really distributed as $F$. Instead, the distribution has higher tails, and thus results are more liberal than they should be when interpreted using an ordinary $F$ table. The reason the distribution is not an $F$ distribution can be seen if you consider many replications of experiments like the one we have just analyzed. Over the series of experiments, we would expect the value of an ordinary $F$ statistic to have an $F$ distribution under the null hypothesis. But when we make corrections, it is not just chance variations in the sums of squares which effect our computed $F$ statistic, but also variations in the estimate of reliability we use in the formula. Thus, the variability in the corrected $F$ is increased because of the variability in our estimates of reliability. Just how much effect this has is unknown, since often the variance of the reliability estimates is unknown. As a simplifying assumption, for the analyses in this report the tabled $F$ values will be used. Thus, unreliability in estimating reliability is not taken into account.

In this example, we can see that adjusting the sum of squares has changed the result from being significant to being nonsignificant. An estimate of the treatment effects can be made by computing the difference in posttest scores...
between the two groups adjusted for the pretest difference:

\[ \Delta Y - B^* \Delta X \]

where \( B^* = \frac{B}{B} \) is the adjusted beta weight.

For the second grade students taking the Metropolitan test,

\[
\begin{align*}
\bar{Y}_0 &= 30.74 & \bar{Y}_1 &= 23.31 & \Delta Y &= \bar{Y}_0 - \bar{Y}_1 &= 7.43 \\
\bar{X}_0 &= 20.60 & \bar{X}_1 &= 17.36 & \Delta X &= \bar{X}_0 - \bar{X}_1 &= 3.24
\end{align*}
\]

where \( \bar{Y} \) indicates a posttest mean, \( \bar{X} \) indicates a pretest mean, a subscript of 0 indicates the control group, and a subscript of 1 indicates the experimental group.

Thus, for this example, the corrected difference between the control and experimental groups assuming perfect reliability would be

\[
(30.74 - 23.31) - (\frac{31367.12}{30982.87})(20.60 - 17.36)
\]

\[
= 7.43 - (1.0124)(3.24) = 4.15
\]

If the reliability estimate is .60, then the estimate of treatment effects becomes

\[
7.43 - (1.687)(3.24) = 1.96
\]

Therefore, making the correction gives a much smaller estimate of the treatment effect, which had been significant when perfect reliability was assumed, but is not significant when the correction is made.

The DeGracie-Fuller approach has the same aim as the Porter approach to ANCOVA, the objective being to correct for unreliability in the covariate. Their method is more complicated, both theoretically and computationally,
but the process is nonetheless straightforward and can be done using the same sums of squares table already constructed for an ordinary ANCOVA. The example presented here has been computed using the same data and reliability estimate as for the Porter correction, so that the two results can be compared.

The first step in the procedure is to compute the following quantities:

\[
\sigma_E^2 = (1 - R_X) \frac{SS_{W_X}}{df_W} = (1 - .60)(30982.87/956) = 12.96
\]

\[
\sigma_T^2 = R_X \frac{SS_{W_X}}{df_W} = (.60)(30982.87/956) = 19.45
\]

\[
\sigma_{XY}^2 = \frac{SS_{WXY}}{df_W} = 31367.12/956 = 32.81
\]

\[
\sigma^4 = \frac{(N-1)/(N-3)}{\sigma_E^2} = (957/959)(12.96)^2 = 167.61
\]

Now we compute the new estimate of $\beta$ using the DeGracie-Fuller approach

\[
\hat{\beta} = \frac{\sigma_{XY}}{\sigma_T^2 + (2/(N-2))\sigma_E^2 + (2\sigma^4/\sigma_T^2)}
\]

\[
= 32.81/(19.45 + (2/956)(12.96 + 2(167.61)/19.45)) = 1.68
\]

In this case, this estimate is the same as that given by
the Porter method, which was $\hat{\beta}^* = \hat{\beta}/R = SS_{WXY}/(SS_{WX}(R)) = 1.68$.

Now the within-groups variance is computed:

$$\sigma^2 = (1/(N-3))(SS_{WY} - 2\hat{\beta}(SS_{WXY}) + \hat{\beta}^2(SS_{WX}))$$
$$= (1/955) (79193.27 - 2(1.68)(31367.12) + (1.68)^2(30982.87))$$
$$= 64.132$$

Now the quantity gamma is computed:

$$\gamma = (1/(N-2))(1/(\sigma^2_{\hat{\beta}}))(SS_{WY}/(N-2)) + ((\hat{\beta}^2\sigma^2_{\hat{\beta}})/\sigma^2_{\hat{\beta}})(1+2(N-3)/(N-1))$$
$$= (1/956)(1/(19.45)^2)((30982.87)/(956) + ((1.68)^2(167.61)/64.132)(1+$$
$$2(955)/957)$$
$$= .0001507$$

Next the between sum of squares is derived:

$$SS_B = SS_{BY} - 2\hat{\beta}(SS_{BYX}) + \hat{\beta}^2(SS_{BX})$$
$$- (\gamma/(1+\gamma(SS_{BX})))(SS_{BYX} - \hat{\beta}(SS_{BX}))^2$$
$$= 2216.23 - 2(1.68)(968.63) + (1.68)^2(422.91)$$
$$- (.0001507/(1+.0001507)(422.91))(968.63 - (1.68)(422.91))^2$$
$$= 145.81$$

The F ratio, with 1 and N-3 degrees of freedom, is computed as the ratio of the between mean square (which equals the between sum of squares in this two group case)
and the within variance:

\[ F = \frac{SS_g}{\sigma_w^2} = \frac{145.81}{64.132} = 2.27 \text{ with } 1,955 \text{ df.} \]

Both estimates took what at first glance appeared to be a significant harmful effect and showed that when corrections are made, the effect is not significant.

It should be noted that using internal consistency estimates in formulas for corrections may give wrong answers under most circumstances. Consider what would happen if there were reliable variance in the covariate which was not contained in the dependent variable. This might happen, for example, in a case where an intelligence test was used as a covariate. If the intelligence test was a multiple-choice test, then part of the reliable variance in performance on the test will be due to intelligence, but part will be due to ability to do well on multiple-choice tests. What we would like to do is to estimate the proportion of variance due to intelligence, which is shared with the dependent variable. But the reliable variance due to performing either well or poorly on multiple-choice tests would not be contained in the dependent variable unless it too was a multiple-choice test. Thus, the internal consistency is an upper bound on the reliability estimate we should use in making corrections.

The situation becomes more complicated when you consider the possibility that the experimental and control
groups differ on more than one factor at the start of the research. In this case, the covariate must contain these same factors, and in the same proportion as the dependent variable. If, for example, 30% of the variance of the dependent variable is determined by quantitative ability, and 20% is determined by verbal ability, then the covariate (or scale made by combining the covariates) must contain proportional amounts of variance due to those sources. Otherwise, when you correct for verbal ability you would either overcorrect or undercorrect for quantitative ability, and vice versa.

Another consideration is that the form of the relationship between the underlying factor and the dependent variable must be the same as the form of the relationship between that factor and the covariate. Otherwise, the effect on the dependent variable of a change in the covariate could differ for the experimental and control groups.

At the lower end of possible values to use in the formulas for correction is the correlation between the covariate and the dependent variable. If the dependent variable and covariate are parallel measures, then this will prove to be a good estimate to use. If, on the other hand, there is reliable variance in the dependent variable which is not measured by the covariate, then the correlation will
be too low, and overcorrection will result. Also, if errors in the dependent variable and the covariate are not identically distributed, then either overcorrection or undercorrection could result. In the research reported here, both internal consistency estimate and correlations between the covariate and dependent variable were used to provide estimates of the possible range of treatment effects.

An alternative to the pretest

Some of the common methods of analysis discussed in this paper seem to depend on the availability of a pretest; that is, a test which is either identical to or parallel to the posttest. For example, gain scores and standardized gain scores cannot be computed if the pretest and posttest are not parallel. Techniques which depend on pretests in a less obvious way are those which require reliability estimates to correct computations, such as the methods advocated by Porter and DeGracie and Fuller. Even though there is some controversy as to what type of reliability measure is appropriate for these methods, both internal consistency and test-retest estimates are usually available, and sometimes do not differ much. The use of other covariates presents several problems: How do you measure the internal consistency of, for example, mother's education? Is it more appropriate to use the
covariate-posttest correlation as the reliability estimate (similar to test-retest reliability), or would it be better to use a measure of internal consistency? Would the results be valid for the possible low values of covariate-posttest correlation which might occur in these situations?

In the analyses of the data in this report one such alternative to a pretest was used. Since studies of compensatory education often contain information on indicators of socioeconomic status which are measured in fairly consistent ways (for example, mother's education, father's education, occupational level, and family income), these variables represented obvious measures to try. The variables were first standardized, and then added together to give a socioeconomic status (SES) scale. One advantage of using such a scale is that an estimate of internal consistency (Cronbach's alpha was used in this instance) can be easily computed. This will serve as an upper bound for possible reliability values in calculating ANCOVAs corrected for unreliability.

This method of simply adding standardized scores is an obvious one to try, but is there any evidence that it is a good method? The answer is contained in an article by Wainer (1976) on estimating coefficients in linear models. There are several reasons why using equal weights would be handy even if the results were not optimum: they are easy
to estimate, do not use any degrees of freedom, are insensitive to outliers (unlike least squares), and are not disturbed by nonnormality as much as least squares estimates derived from regressing the dependent variable on multiple covariates. Wainer states that there is in fact almost no loss in accuracy when least squares weights are replaced by equal weights, and that on cross-validation, equal weights are more robust. This is because equal weights do not capitalize on chance (as do least squares weights), and they are relatively insensitive to outliers. As a final argument, Wainer cites standard practices in test construction as following this method, and notes that most attempts to use differential scoring of tests have not improved much on equal weighting.

**Factor analysis methods**

**Combining factor analysis and regression.** Since factor analysis was developed to define underlying traits ('true scores'), some researchers have attempted to solve the problem of errors in variables by using methods of factor analysis, either by itself or in combination with multiple regression methods.

The problems of errors in independent variables and high intercorrelations among independent variables have been taken more seriously by economists than psychologists, since psychologists have been able for the most part to work
within an analysis of variance framework and have randomly assigned subjects to groups. Scott (1966) saw factor analysis as a way to estimate the underlying structure of economic problems and thus eliminate biases due to errors in independent variables.

Two approaches to using factor analysis will be examined in this study. One method involves combining factor analysis and multiple regression. This method was outlined by Scott (1966) and applied to a problem in economics. Scott listed two slightly different ways for estimating regression coefficients from factor analysis, one of which was later used by Jurs (1971) in a doctoral thesis examining the effects of compensatory education, while the other was later extended by Lawley and Maxwell (1973), who also found unbiased estimators of regression coefficients through factor analysis.

A different approach involving factor analysis is to try to fit different models of the underlying processes and test them using factor analytic technique. The technical tools to do this have been developed by Joreskog (1969; Joreskog, Gruvaeus, & van Thillo, Note 3) who has also developed several computer programs (EMLFA, ACOVS, LISREL) to perform the calculations necessary for this model testing. The procedures outlined by Scott can be expressed in terms of operations on the basic matrices in the factor
analysis model, $Z = AF + U$, where

$Z$ is an $(n \times 1)$ vector of standardized variables,

$A$ is an $(n \times m)$ matrix of factor loadings,

$F$ is an $(m \times 1)$ vector of factors,

$U$ is an $(n \times 1)$ vector of error terms (uniqueness).

It is assumed in the model that $E(U) = E(F) = 0$, $E(UU') = V$, which is a diagonal matrix, and $E(FF') = I$, the identity matrix. Thus, the factors are uncorrelated, with the variance of each equal to 1, the errors are uncorrelated, and the errors and factors have means of 0. Furthermore, the errors are assumed to be independent of the factors, and both errors and factors have multivariate normal distributions. It follows that the expected covariance matrix $R = AA' + V$. The diagonal elements of $AA'$ are called the communalities and the elements of $V$ are the specific variances of the variables.

After the matrices $A$ and $V$ are obtained by factor analysis, the factors can be obtained by solving any of the following equations listed by Scott:

1. $F = (A'A)^{-1}A'Z$
2. $F = A'R^{-1}Z$
3. $F = (I + A'V^{-1}A)^{-1}A'V^{-1}Z$

Equation (2) is equivalent exactly to equation (3), and although (3) looks much more complicated it is often easier to use because the computations involved are simpler than
inverting the matrix $R$ in equation (2), which is especially important if the calculations are not done on a computer.

Scott notes that equation (1) disregards the problem of errors in variables, since $Z$ is used instead of $Z - U$. This can be seen by noting that $Z = AF + U$, which, upon premultiplication by $(A' A)^{-1} A'$ and collecting terms gives $(A' A)^{-1} A' Z = F + (A' A)^{-1} A' U$. Subtraction and applying the distributive law yields $F = (A' A)^{-1} A' (Z - U)$, which is the same as equation (1) except it contains $Z - U$ instead of $Z$. Equation (1) is analogous to the least squares solution of ordinary linear regression. The model for regression is $Y = XE + E$, and the least squares estimator of $B$, denoted here by $B^*$, is calculated by the equation $E^* = (X' X)^{-1} X' Y$.

After presenting these two models, Scott derives estimators for predicting one variable from the others, comments on some statistical tests which he feels are appropriate (although he notes that he has no proof), and applies the model to a problem in predicting expenditures on building investment, comparing the factor analysis prediction equations with ordinary least squares regression. His mathematical treatment will not be presented here because it is much more cumbersome and limited than the matrix treatment given by Lawley and Maxwell (1973).

In Scott's comparison of the two methods he concludes that the factor analysis method was better than the least
squares method because the signs (+ or -) of the coefficients for the variables were more consistent with theoretical expectations in the factor analysis model. Scott also notes that the factor analysis method gives lower estimates of proportion of variance accounted for ($R^2$) than traditional regression methods, but says that this is to be expected since the factor analysis model takes into account errors in variables. (In view of this, it is puzzling that Scott, in his empirical analysis, uses equation (1) because it "consistently gives the highest multiple coefficient of determination $R^2$."

Lawley and Maxwell (1973) examined the implications and methodology of combining factor analysis and regression in more detail than Scott. For the purposes of this paper, however, only two aspects of their approach will be presented in detail. The notation they use is different than that of Scott, so it will be changed slightly to conform to that already used in previous sections of this paper.

First, Lawley and Maxwell partition the matrices of factor loadings and covariances:

$$A = \begin{bmatrix} a_1 \\ A_2 \end{bmatrix} \quad R = \begin{bmatrix} S_{11} & a_1' A_2' \\ A_2 a_1 & R_{22} \end{bmatrix}$$

In these matrices, $a_1$ represents the factor loadings on
variable 1, which is considered the dependent variable to be predicted from the other variables. \( A_z \) represents the remainder of the factor loadings. The element \( s_{ii} = a_i'a_i + u_i \) is the variance of \( x_i \), the dependent variable, and \( R_{zz} = A_z'A_z + U_z \) is the covariance matrix of \( x_z \), the vector of independent variables.

When there are errors in all of the variables, Lawley and Maxwell state that the least squares estimator gives the following prediction equation:

\[
x_1^* = a_1'(I + A_z'R_{zz}^{-1}A_z)^{-1}A_z'U_z^{-1}x_z
\]

Next, Lawley and Maxwell derive estimators which are least squares subject to the constraint of being unbiased:

\[
x_1^* = a_1'(A_z'R_{zz}^{-1}A_z)^{-1}A_z'U_z^{-1}x_z
\]

The multiple correlation between the dependent variable \( x_i \) and the predicted dependent variable \( x_i^* \) is given for the first estimator by the equation

\[
P^2 = (a_1'a_1 - a_1'(I + A_z'R_{zz}^{-1}A_z)^{-1}a_1)/s_{ii}
\]

and the equation for the unbiased estimators is given by

\[
P^2 = (a_1'a_1)^2/(s_{ii} a_1'(I + A_z'R_{zz}^{-1}A_z)a_1), \text{ where } P \text{ is the multiple correlation coefficient.}
\] Lawley and Maxwell state that for the unbiased estimators, the multiple correlation will always be less than or equal to that obtained from the biased estimators. If there is only one factor, then the two estimates will give exactly the same multiple correlation.
What would we expect the results of combining factor analysis and regression in the manner suggested by Scott and by Lawley and Maxwell to be in the case where there is biased assignment to treatment? If the treatment were truly ineffective, would using a combination of factor analysis and regression discover this? The answer is no, it would not unless the covariates were sufficient to explain all reliable within group variation in the dependent variable. Why is this so? If the assignment to treatment is biased, then probably knowing which treatment group a subject is in adds knowledge of his underlying ability. In many cases, whether a person is assigned to treatment is not based solely on variables which are used as covariates, but on judgments which probably do add some new information. Thus, even if there were no treatment given, the covariate defined by assignment to treatment groups would be a variable which would help predict the dependent variable. Thus, we would not expect the coefficient of treatment condition to be zero in most real-life studies even if the treatment were ineffective. The conclusion is that we cannot use the Lawley-Maxwell procedure in the same way as we use Porter or Lord's procedure in the case of using a pretest as a covariate. This will be demonstrated in the analysis of data constructed by computer simulations from a one-factor model.
Fitting factor models. Brewer, Campbell, and Crano (1970) have suggested using factor analysis in situations similar to our original problem of estimating treatment effects. In many cases, investigators have tried to establish the effect of a variable when other variables are 'controlled' for by using the technique of partial correlation. As we have noted, this technique does not give correct estimates if there are errors in the variables. Suggestions have been made to correct correlations for attenuation before doing a partial correlation (e.g. Cohen & Cohen, 1975). But sometimes reliability in the traditional sense is not what we want to correct for, but rather irrelevance; that is, variance unrelated to the dependent variable. Another way of stating this is that "all variables which might affect the dependent variable are included in the regression equation or are uncorrelated with the variables which are included (Darlington, 1968)." For example, the variable of race may be measured with almost perfect reliability, but correcting an analysis of variance using race as a covariate would not remove all of the bias of assignment to treatment. Thus, correction for attenuation is usually not possible because measures of relevance are not available. For this reason, Brewer et al. suggested testing a one-factor model first, and only if this model is rejected would any significance be attached to the
partial correlation. Another possibility would be to factor analyze the data without restrictions and use the communalities as estimates of relevance, and use these estimates in formulas for computing partial correlation with correction for attenuation.

If, in fact, a particular treatment is ineffective and one underlying factor is sufficient to explain the relationship between the posttest, treatment assignment, and the covariates, then this should be detectable through factor analysis. In particular, Joreskog, as mentioned earlier, has developed computer programs which enable the testing of any linear structural model assumed to be responsible for a pattern of correlations. As an example of how these programs might be used, consider the problem of detecting which of the following two models (if either) is correct:

(1) only one factor underlies all of the relationships between the dependent variable, treatment group, and the covariates, or

(2) two factors are necessary to account for the relationship. One is a general factor which influences all the covariates and assignment to treatment (thus producing biased assignment), and the other factor is a treatment factor, which would not show up in the relationship of treatment to the independent variables. Thus, the second
factor would be hypothesized to have nonzero loadings only on the dependent variable and the dummy variable for assignment to treatment condition.

To see which model is correct, programs would be run with appropriate restrictions to fit each model. A chi-square statistic is computed by the program to tell if each model provides a reasonable fit to the data. If the chi-square for both models is large, then a more complicated model is necessary. Perhaps then models with two general factors might be tried, with one model containing only the two general factors (which might in psychological terms correspond to intelligence and economic advantage in most compensatory education studies), and another model containing an additional specific factor which would load only on the dependent variable and treatment group dummy variable.

If the chi-square for both models is small, then the first would be accepted, since it is simpler. This result would indicate that one general factor (probably interpreted as either intelligence or socioeconomic status in enrichment studies) was sufficient to account for the relationships in the data, and that the second specific factor which tested for a treatment effect was unnecessary. If the chi-square for model 1 is large enough to reject that model, while that for model 2 is small, then it would be concluded that the
treatment has an effect. The direction of the effect could then be determined by comparing the sign on the loading for treatment group with the sign of the loading for the dependent variable in the specific factor. An example of how this technique could be applied is contained in the section describing the analysis of simulated data.

There are two potential problems with this method, only one of which is solvable. The solvable problem is that when the sample size is large, almost any model will fail the chi-square test. This is the same as in any other test of a null hypothesis: if you get a large enough sample you will reject it. Thus, a measure which is less dependent on sample size would be preferred. One solution is to use the size of the residuals, which are found in the matrix $K = S - (AA' + V)$. This tells how close the correlations predicted by the factor model come to the actual correlations between variables. If these are 'small' (this is where subjectivity creeps in), then the model is said to fit.

In some situations, though, one or two factors will not explain the variance, and neither the chi-square test nor the examination of residuals will allow a simple model to be accepted. This problem is unsolvable: if the solution is so complex factorially that the factors cannot be interpreted easily, then the assessment of treatment effects by this method is impossible.
Summary

The existence of regression artifacts in nonexperimental data makes them very difficult to analyze. Although many techniques have been suggested for solving this problem, some are insufficient, and all can fail if the data does not "behave." Corrections can be made for unreliability in an analysis of covariance, but only if the data meet the assumptions of equal slopes. Even then, there is an argument about what estimate of reliability is appropriate. Regression on factor scores does not adequately correct for unreliability and is thus useless in this situation. Fitting factor models to data can be useful if the factor structure of the data is simple, but it remains to be seen if real data fit the model. Many of these methods are improvements over current practice, but none would meet universal approval from researchers in this field.
Simulated Data

Method

The simulated data set was constructed according to a one-factor model presented in Campbell and Boruch (1976). The model is that POSTTEST and PRETEST each share 80% of their variance with the hypothesized underlying factor, and TREATMENT, COVARIATE 1, COVARIATE 2, and COVARIATE 3 all share 50% of their variance with the common underlying factor. A sample consisting of 100 cases each from an 'experimental' and 'control' group were generated by computer sampling from a multivariate normal distribution. The correlation coefficients for this model are given in Table 1. The population values derived from the model are below the diagonal, and the sample values are above the diagonal.

The variable TREATMENT is a dummy variable signifying assignment to the experimental or control group. Since this variable is dichotomous, its correlations with the other covariates are lower even though it shares as much of the underlying variance. This simulation represents a situation where treatment was given to the neediest, since treated subjects had scored lowest on the pretest. Thus, assignment to treatment is strongly biased, but there is no effect of treatment. This can be easily seen by noting that the treatment-pretest correlation is equal to the
treatment-posttest correlation. But the methods we will test here are used in situations where no pretest is available, so the pretest score will not be used in these analyses.

To demonstrate what the factor analysis-regression technique accomplishes, the following analyses will be compared: the population values of the preceding model will be analyzed using ordinary multiple regression, factor analysis regression, and unbiased factor analysis regression; and the sample data generated from the model (N=200) will be analyzed using the same three methods. For the population values, the coefficients for each of the three covariates will be equal for any method used. However, the sample values will vary, so that, for example, the beta weight for Covariate 1 will not equal that for Covariate 2, and so on. What should happen for the sample data is that the coefficients of the covariates should vary more when ordinary regression is used than when factor analysis regression is used. That is, factor analysis regression estimates should be more stable than ordinary least squares estimates.

Results

The standardized regression coefficients (beta weights) and the $R^2$ values for each of the three types of analysis of each of the two correlation matrices are in Table 2. Also
included in this table are estimates of the regression weights and $R^2$ values when the treatment variable (Trt) is included in the analysis.

First, note that there is a fairly wide range of estimates of the beta weights for the ordinary regression method, while the range for the least squares factor analysis regression method is smaller. Thus, the factor analysis regression estimates are more stable. Next, notice that the coefficient for treatment is fairly large regardless of which type of regression estimate is used. In fact, even though the $R^2$ change is small, it is significant, because the number of subjects is fairly large.

Thus, the factor analysis regression method accomplishes the purpose of providing more stable estimates, but does not provide unbiased estimates of treatment effects. It correctly provides the information that knowing treatment assignment helps predict test score, since treatment assignment is a variable which contains information about the students' abilities not contained in the covariates.

The simulated data was also used to test the potential usefulness of fitting factor models to data. The data described were analyzed both as they were (with no treatment effect) as well as with three different values of treatment effect added to the posttest scores of the experimental
group. These treatment effects correspond to small, medium, and large size effects as described in Cohen's (1969) book on power analysis. Two factor analyses were done on each of the four correlation matrices (zero, small, medium, and large treatment effects). One analysis was a test of a one factor fit, and the other allowed a second factor which was restricted (by the computer program) to allow nonzero loadings only on the posttest and the dummy variable indicating assignment to treatment condition.

For all four matrices, results are in Table 3. For both one factor and two factor models, chi-square values and associated probability levels are listed. Also, for the two factor model the loadings on the second (specific) factor are presented for the only variables allowed to load on that factor—posttest score and treatment condition.

Note that the chi-square test successfully found medium and large treatment effects, since the one factor models were rejected while the two factor models were accepted. Although the small treatment effect was not detected by the chi-square test, the loadings on the second factor were both positive. For medium and large size effects, the loadings were fairly large, and in all cases the two factor model fit
the data. Thus, this method of analysis has potential for detecting treatment effects if they are large enough (it does not have enough power to detect very small effects), and if a simple model will fit the data.
Real Data

Method

Following the procedures outlined in the introduction, two data sets were analyzed using all three methods: multiple regression, factor analysis regression, and testing hypothesized factor structures. One data set was from the evaluation of Project Head Start, whose purpose was to give physical, psychological, and cognitive aid to deprived children just before they entered formal schooling. The other data set analyzed is from the early programs funded by Title 1 of the Elementary and Secondary Education Act of 1965. The Title 1 programs had very broad goals, so only a portion of the subjects' records were analyzed. The selection was based on whether the treated childrens' neighborhood had received remedial reading programs (cognitive skills development) rather than just psychological or physical help (hot lunches, physical examinations).

The following section describes the data which were analyzed in this study. These data sets were both obtained from the results of projects sponsored by the Office of Educational Opportunity to evaluate nationwide programs of compensatory education. Only part of each data set was analyzed.
Head Start. Project Head Start was a national attempt at preschool education for disadvantaged children. Its purpose was to prepare these children for formal education by improving their physical, emotional, mental, and social skills as well as cognitive skills. The program was implemented through a variety of techniques, but there was no standard program administered throughout the country. This not only makes the pinpointing of causes of effects more difficult, but also makes it unlikely that there would be very large effects, since the program was so diffuse.

The original analysis of the data was done by Westinghouse Learning Corporation (Cicirelli, 1969). The purpose of the analysis was to assess the effect of treatment on cognitive skills. Since the administration of the program had been oriented toward providing service instead of evaluation, there was no pretest and the control group had to be selected after the treatment had been administered.

In order to reduce differences between the experimental and the control groups, but at the expense of generalizability, the investigators decided that all subjects had to satisfy four conditions to be included in the analysis: continuity of residence in the target area, eligibility for Head Start (all subjects were eligible to have participated in Head Start), attendance of the same
school system, and no other experience with Head Start except for the program being evaluated. One hundred and four centers were finally used from a random sample of the nearly 13,000 Head Start centers existing in 1966-67. However, 225 centers had to be screened before the researchers found 104 which were willing and able (in terms of providing enough subjects) to participate. Within each target area eight former Head Start participants and eight eligible nonparticipants each from the first, second, and third grades were chosen. The subjects who had Head Start participation were randomly selected, while the control subjects were chosen randomly subject to the constraint of being matched to the experimental subjects on age, sex, kindergarten experience, and racial/ethnic characteristics. Since children were matched, the $F$-statistics are evaluated using the number of pairs of subjects minus one as the total degrees of freedom, rather than the number of subjects minus one. This makes the $F$ tests somewhat conservative, but it only affects borderline cases since the number of subjects (and thus, the number of pairs of subjects) is large.

Since so many centers had to be tried before the final sample was obtained, the investigators made an attempt to determine if there was any bias in the final sample. An interview questionnaire was sent to all of the 225 centers, and the answers from those who had become a part of the
study were compared with those which hadn't. Only 54 of the 121 nonparticipating centers returned the questionnaire. On five of the 32 items the participating group differed from the nonparticipating group at the .05 level. Thus, the sampling issue remains unsettled, since the nonresponding nonparticipant group may be different from the responding nonparticipant group.

Although the original investigators measured affective outcomes in addition to cognitive outcomes, only the latter were analyzed in this study. In particular, only the results for those subjects who were in the first grade, had summer Head Start experience (in contrast to full year Head Start), and who had both parents living with them were included in the analysis. Two measures of cognitive ability were used. One was the Metropolitan Readiness Test (MRT), which has six subtests: word meaning, listening, matching, alphabet, numbers, and copying. The authors of the test state that the total score (the sum of the subtest scores) should be used instead of trying to analyze each subtest score, because the total score has a reliability of .91, while the reliability of the subtests ranges from .50 to .86. The other measure of cognitive skills used was the Illinois Test of Psycholinguistic Abilities (ITPA). The object of the ITPA is to aid in the diagnosis of specific abilities and disabilities and to guide in the
administration of remedial work. The 10 subtests of the ITFA are auditory reception, visual reception, auditory-vocal association, visual-motor association, verbal expression, manual expression, grammatic closure, visual closure, auditory sequential memory, and visual sequential memory.

The results of the original analysis led the researchers to conclude that participation in summer Head Start programs had no effect on ITFA scores, and may have had a negative effect on some subtest scores on the KRT. Since first graders (especially whites) who only had summer exposure to Head Start seemed to fare the worst, only their data was analyzed.

Independent variables included in these reanalyses were those which might reflect possible advantage for control subjects in parents' social, educational, or economic status. The variables used were:

1. Total annual family income.
2. Mother's educational level. This was recoded from the original data as 5 if the mother had over 12 years of education, 4 if she had 12 years of education, 3 if she had 10 or 11 years of education, 2 if she had 7-9 years of education, and 1 if she had 0-6 years of education.
3. Father's educational level. This was coded the same way as mother's educational level.
(4) Father's occupational level. This was recoded from the original data as 5 if the father was in a professional or managerial position, 4 if he was a clerical worker, 3 if he was in a skilled occupation, 2 if in a semi-skilled position, and 1 if in an unskilled occupation or unemployed.

(5) Race.

Title 1. Title 1 of the Elementary and Secondary Education Act of 1965 had two objectives: (a) to provide equal educational opportunity for all social and ethnic groups, and (b) to reduce deficiencies in educational attainment that were associated with social class membership. Local administrators of the program were allowed to decide how to implement programs to achieve these goals; there were no overall plans given in the law. Over 5,000,000 children participated in Title 1 programs. In general, the student's teacher made the decisions about who would participate in Title 1 programs, but any student in a school which was eligible to receive Title 1 funds could participate. Since over 90% of the nation's schools were eligible for funds, many children who did not need special remedial help received it anyway. In fact, about 36% of the participants in remedial reading programs who had taken the Metropolitan achievement test were reading above grade level, and 12% of those who had taken the Stanford achievement test were reading above grade level. While only
about 36% of those who were in need of help (reading below grade level) received it, a comparison of participants versus nonparticipants show that, in general, more of the needy were exposed to the program than those without need.

The average amount spent per poor child nationally from Title 1 funds was only about $157 per year. Rich school districts got more per poor child than poorer districts, since benefits were linked to school district expenditures.

The extent of benefits given through Title 1 funding is not overly impressive. About 90% of the respondents to the survey on usage of funds reported using the funds for providing either teacher aides (usually parents of some of the pupils) or consultants. Only about 70% of pupils receiving aid under Title 1 received more than 100 hours of remedial instruction a year (less than 3 hours per week).

Measuring the impact of Title 1 programs on the cognitive skills of students was difficult since only about 7-1/2% of the participating students had both pretest and posttest scores. Most of these students were from large urban districts where such standardized testing is more likely to be mandatory. Only in the area of reading instruction were enough subjects available to do an analysis. The data analyzed here is for second graders only, although the original study included fourth and sixth graders also.
In the original analysis (Glass, Note 2) subjects were divided into three groups: nonparticipants, who had no remedial training through Title 1 programs; participants in only one Title 1 program; and participants in two or more programs. For both pretest and posttest scores, discrepancy scores were calculated by subtracting from the grade level achieved on the test the actual grade level of the student. Thus, if a second grader took a test in October and received a grade equivalent score of 2.5, his discrepancy score would be 2.5 - 2.1 = .4 years. The pretest discrepancy scores showed that nonparticipants scored higher than those participating in one Title 1 program, who in turn scored higher than those participating in two or more programs. Thus, more remedial treatment was given to those who appeared to need it more. Data analysts must be careful of regression artifacts in this type of situation, since the control subjects would be expected to regress toward a different mean than the experimental subjects.

In comparing the posttest discrepancy scores with pretest discrepancy scores for the three groups, Glass (Note 2, p. 116) concluded that "...analyses of reading achievement gain scores show nonparticipants to have made larger gains than participants in either one or in two or more disadvantaged reading programs. 'Participants' of either type tended to lost ground during the course of the
school year: "nonparticipants did not."

As was the case with the Head Start data set, the independent variables used in this analysis were selected to reflect possible social, educational, or economic differences between families of participants and nonparticipants. The availability of pretest scores in this instance allows us to compare the techniques being investigated here with other techniques which have been advocated.

(1) Pretest score. Pupils took either the Stanford Primary I achievement test or the Metropolitan Primary I achievement test. Scores were reported in grade equivalent terms for all pupils and in percentiles for some. Only the grade equivalent scores were used.

(2) Posttest score. Pupils took either the Stanford Primary II (if they had taken the Stanford pretest), or the Metropolitan Primary II achievement test (if they had taken the Metropolitan as a pretest). Note that because the posttest was not the same as the pretest, gain scores of any kind may be misleading because of norming differences. Furthermore, scores on the Stanford are probably not comparable with scores on the Metropolitan, so the results were analyzed separately for students who took each test.

(3) Participation in Title 1 programs. In order to obtain a stronger test of possible treatment effects, only pupils
who had been either nonparticipants (control group) or who had participated in two or more programs (experimental group) were included in the analysis. Those who had participated in only one program were not used in the analysis.

(4) Family income. This was estimated by the teachers with the help of school records, and thus may not be very accurate. Income was estimated as being either under $3000, $3000-$4500, $4500-$6000, $6000-$7500, $7500-$9000, or over $9000.

(5) Occupation of head of household. This was coded as 6 for professionals, 5 for those in technical occupations, 4 for owner-managers, 3 for skilled laborers, 2 for semi-skilled, and 1 for unskilled.

(6) Educational level of head of household. This was coded as 6 for completion of college, 5 for some post high school education, 4 for high school graduation, 3 for some high school, 2 for grade school education, and 1 for little or no formal education.

(7) Educational level of mother. This was coded the same way as educational level of head of household.

Results

**Title 1 Metropolitan test.** When the actual pretest was used as a covariate, a significant effect was found in the uncorrected ANCOVA \( F(1,955) = 13.69, p < .001 \). Table 4
contains the means and variances for all of the data analyzed. The adjusted mean difference between groups was -4.14, meaning that the experimental group averaged 4.14 points lower on the posttest than the control group, 'corrected' for differences on the covariate. Table 5 shows that F values and corrected mean differences for a range of possible reliability values. When the reliability is assumed to go down (reading down the columns of the table) the F value decreases rapidly. It is significant (p<.05) for all reliability values greater than or equal to .65, and is nonsignificant for values of less than .65. The test-retest correlations were .635 for the control group and .5157 for the experimental group. Thus, a case can be made for borderline significance using the value from the control group, which could not have been affected by the treatment. No estimate of internal consistency for either group was available, but it can probably be assumed to be over .80, which would imply a significant negative effect of Title 1.

Next consider what happens when we use as the covariate the composite variable consisting of the sum of four variables, the four variables being the standardized SES indicators mentioned previously (mother's education, father's education, father's occupation, and family income).
Cronbach's coefficient alpha was computed for each group. For the control group, alpha was .433, while for the experimental group it was -.031. Since the proposed covariate was totally unreliable for the experimental group, the results would be misleading. These results were computed anyway as a demonstration of the possible effects of violating assumptions: the effect is negative and significant for all values of reliability ($F(1,955) > 29$, $p < .001$).

**Title 1 Stanford test.** When the pretest was used as a covariate, the effect was again significantly negative when uncorrected ($F(1,862) = 4.19$, $p < .05$). The results of making the corrections are presented in Table 6. As the estimate of reliability is lowered, the value of $F$ quickly becomes nonsignificant and reaches a value of zero for a reliability estimate of .70. As the reliability estimate drops further, the $F$ values computed by the Porter method increase, this time indicating a positive treatment effect. Since the test-retest correlation was .7075 for the control group and .6748 for the experimental group, it would appear that there is no real difference between the groups.

When the scale of the four standardized SES variables was used as covariate the results were better than for those who had taken the Metropolitan test. Cronbach's alpha was .525 for the control group, and .537 for the experimental
group. When no correction is made for unreliability, the results again appear to be significant and negative ($F(1, 862) = 12.02, p < .001$). However, as the reliability estimates decrease (see Table 7) so do the values of $F$, until they are no longer significantly negative when the reliability is less than .55. The $F$ value reaches 0 when the reliability estimate is about .35. The estimates of Cronbach's alpha provide a reasonable estimate of the upper bound of possible reliability values, while the correlations between the covariate and the dependent variable (.3943 for the control group, .3420 for the experimental group) provide reasonable lower bounds for possible reliability values. Thus, the effect is estimated to be nonsignificant throughout most of the region of possible reliability values, reaching borderline significance at the $p = .05$ level only at the upper limit of possible values for the Porter correction. For the DeGracie-Fuller corrections, the region of nonsignificance is even larger, and includes all of our reasonable values for reliability. Keep in mind that the tabled probability values are too liberal by an unknown amount. Thus, looking at a range of possible values of reliability allows us to better determine how stable our results would be if our best guess about the reliability turns out to be wrong. In this particular case, our results are stable over a wide range of possible reliability
estimates.

**Head Start Blacks.** For the ITPA the ANCOVA provides misleading results because the correlation between the covariate and posttest is near zero for the experimental group. The correlation for the experimental group was .0476, while for the control group it was .4352. Thus, the results (which show consistent superiority for those in Head Start) are invalid. For all values of reliability, $F$ is greater than 11 ($p<.001$), but as in the Title 1 Metropolitan data, this is because the assumption of equal slopes was violated. In this case, unlike that with the Title 1 Metropolitan data, the low correlation was not due to unreliability in the covariate: for the control group, Cronbach's alpha was .723, and for the experimental group it was .626.

For the MRT scores, the covariate is usable but not very useful. The correlation between covariate and MRT score was .1453 in the control group and .2020 in the experimental group. Thus the use of the covariate changes the results very little. The corrected differences between means were all positive but nonsignificant throughout the range of reliabilities from 1.0 to .30. When the reliability estimate was below .30, the results were significantly positive, indicating a beneficial effect for Head Start.
Head Start whites. The internal consistency of the covariate scale was relatively high for both groups: Cronbach's alpha was .687 in the control group and .650 in the experimental group. The analysis of ITPA means showed that for a large range of possible reliability values (from 1.0 down to .45) the F value was less than 1.0, and thus nonsignificant. For reliability estimates under .45, the effect was zero. For reliability estimates under .45, the effect was positive but did not reach significance until the value of .30 was reached. This is not an unreasonable value, however, since the correlation between the covariate and ITPA score was .3048 for the control group and .2857 for the experimental group.

For MRT means the results are similar. The F statistic is less than 1.0 for reliability values from .45 to 1.0, and reaches significance (indicating a positive effect) for values less than .35 when the Porter correction is used. Again, this is not unreasonable, since the correlations between the covariate and MRT score were .3590 for the control group and .2557 for the experimental group.

Double pretests. At this point, it seems that the situation is reduced to an argument about what type of correction is best, with no one being able to back up the argument with empirical justification for choosing to use internal consistency estimates versus covariate-pretest
correlations. But an out is still available: the double pretest. This is a tactic which was recommended by O'Connor (Note 4) and by Boruch (Note 1), in order to determine what the null hypothesis conditions are for particular quasiexperiments. Suppose you had two pretests, one administered at time T(0) and the other at time T(1), and a posttest administered at time T(2). You could estimate the bias in your null hypothesis conditions by pretending that the pretest at T(1) is really a posttest, and use T(0) scores as your pretest scores. If the method used is really unbiased, the null hypothesis should not be rejected, since at time T(1) there has been no treatment applied, so any differences between groups are pre-existing differences. It should be noted that tests made subsequent to the double pretest analysis are not unbiased, since they are conditional on the results of the double pretest analysis.

The Title 1 data present an (unfortunately infrequent) opportunity to apply this method. There is not only a pretest, but there are the covariates available to make up the SES scale to serve as the other covariate for the double pretest. Two separate kinds of analysis have already been described for the Title 1 data: ANCOVA using a pretest as a covariate, and ANCOVA using a scale of four SES indicators as a covariate. If the uncorrected ANCOVA were appropriate, then its application to the pretest in this manner should
result in failure to reject the null hypothesis, whereas it either the internal consistency or covariate-test reliability are appropriate correction factors, then their use in the appropriate formulas should result in failure to reject the null hypothesis. If none of these methods gives the right answer, then the results of this double pretest should at least indicate the direction and strength of bias in each method.

Using the double pretest method on the Title 1 data provided fairly strong evidence that the correlation between the covariate and the posttest is the best estimate to use in making reliability corrections. For children who took the Stanford test, $F(1,862)=12.7$ when no correction was made. Thus, the treatment was made to look harmful. The effects of making corrections are found in Table 8. The adjusted mean difference between groups was $-2.39$, meaning that the experimental subjects were estimated to have lost about 2 months relative to controls, even after adjusting on the basis of the covariate. Remember that this is the prediction of the pretest score by the covariate and thus the difference should be 0 when proper adjustments are made. When internal consistency (approximately equal to .50) is used as an estimate of reliability, and the appropriate correction formulas are used, the Porter technique gives $F=4.85$, and the DeGracie-Fuller technique gives $F=4.35$. 
Thus, using internal consistency has still undercorrected, making the treatment look harmful. When the correlation between the covariate and dependent variable (about .30) was used as the measure of reliability in the correction formulas, the Porter $F$ was .34 and the DeGracie-Fuller $F$ was .26. Thus, using this correlation gives the correct answer, while using either no correction or using internal consistency leads to undercorrection.

For children who took the Metropolitan test, the slopes for experimental and control subjects differ (as do the estimates of internal consistency) and so the results are misleading, just as they were when the actual posttest was used as the dependent variable. If an ANCOVA is (inappropriately) performed in spite of the violation of assumptions, the treatment is made to look harmful regardless of what kind of correction is attempted. For all values of reliability, the $F$ statistics using either the Porter or DeGracie-Fuller method are about 14 and the estimate of loss for the experimental children relative to the controls is about 3.2 months.

The use of the double pretest has confirmed the earlier hypothesis that the internal consistency estimate of reliability is too high, whereas the covariate-posttest correlation gives correct results. The violation of the assumption of equal slopes led to the same direction and
magnitude of misleading results in the double pretest as in the actual analysis, and the various reliability estimates gave the same results when doing an ANCOVA was actually appropriate. The usefulness of the double pretest has been empirically confirmed by this analysis.

**Fitting factor models.** The attempt to fit factor models to the data proved abortive for the reason mentioned earlier: no simple, easily interpreted models could be fit to the data. As an example of the difficulties encountered, consider the analysis of the Title 1 Metropolitan data. For this analysis, nine variables were used: posttest, assignment to treatment, and seven indicators of socioeconomic status. These seven indicators were selected from a larger set on the basis of a preliminary factor analysis, from which only the variables with the highest communalities were selected for further analysis.

When totally unrestricted models were fit (so that all variables could load on all factors), even a four factor model did not provide good fit. The chi-square for the four factor model was 14.99 with 6 degrees of freedom ($p<.02$). The three factor model, of course, was even worse: the chi-square was 48.50 with 12 degrees of freedom. Thus, even though the variables had been selected to make fitting models easier, there were no simple models which fit the data. When the samples were split into subpopulations (for
example, white-black and kindergarten-nc kindergarten) parts of the samples could be fit to simple models, but other parts could not. Thus, where I started out with one intractable sample, I might end up with two subsamples which were tractable, but still have two more which were not.
Discussion

Two criteria are important in evaluating a technique of analyzing data: Does the technique give the right answer (is it unbiased in the situations where its use is intended)? Is it generally applicable? Many techniques are limited because there are few situations in which they can be applied. For example, gain scores cannot be analyzed if the pretest and posttest are not measured with the same instrument. Often this turns out to prevent errors, since raw gain score analysis would yield biased results in many nonexperimental situations, including the analysis of the Title 1 data.

To see this, we could look, for example, at the possibility of floor effects. An examination of the skewness of the distributions of scores is useful in determining whether there might have been a floor effect. Such an effect would tend to artificially raise pretest scores of the experimental children relative to the controls (assuming that the experimental group actually starts out below the controls), thus making any gain score biased.

One question to consider is whether the distributions were skewed at all. All eight distributions (all combinations of pretest and posttest, controls and experimentals, and Metropolitan or Stanford test) showed significant skew ($p < .01$). All of these were skewed
positively, thus indicating possible floor effects.

A more important question, as far as demonstrating bias, is whether the distributions for controls were less skewed than those for experimentals. In all cases this was true. Thus, the floor effects differentially harm the experimental group, making treatment effects more difficult to demonstrate.

If the control group scores are more reliable than those of the experimental group, a regression artifact could occur, making the control group seem worse than it is. Although the ordering of the reliabilities is consistent with this hypothesis, none of the differences was significant. However, the power of these tests was generally low (under .40), so the significance test results should be taken with a grain of salt.

If the treatment has an effect, the correlation between treatment condition and pretest score should differ from the correlation between treatment condition and posttest score. This did not occur in these data. In absolute terms all effects were small, and no differences were significant, even with more than 370 subjects in all groups. Thus, this method indicates no treatment effects.

Another problem of applicability occurs in ANCOVA and related methods: the problem of unequal slopes in the control and experimental groups. Unfortunately, the problem
often goes unnoticed, since many researchers neglect to check for it. As was seen in the results, if the problem occurs it can lead to biased results which may not be correctable using the techniques outlined here.

It is essential to remember that nonrandom assignment can produce differences on many factors, and for analysis of covariance to be unbiased we must control for all of these factors completely. There are many reasons why our attempts to control for differences may fail. Factors on which the groups initially differ and are contained in the dependent variable may not be measured by the covariate. Even if the covariate does measure the same factors as the dependent variable, they may not be in the covariate in the same proportions as they are in the dependent variable. The covariate(s) may measure the factors fallibly, or measure irrelevant factors. Finally, the relationship between the factors and the covariate may be different than the relationship between the factors and the dependent variable. Only when all of these can be controlled for will the analysis of covariance give unbiased results.

The issue of whether or not a technique gives the 'right' answer is complicated. As indicated in the introduction, some people feel that trying to get the right answer from nonexperimental studies is to be avoided at all cost. Their attitude is that "if
you torture the data long enough, it will confess." (Coase, quoted in Good, 1972) The Title 1 data has been tortured mercilessly, and the Headstart data has at least been abused somewhat, although at this stage it is still difficult to state unequivocally that any confessions have been wrung from the victims.

It is really more accurate to view this study as an attempt to torture methods rather than data. In a certain sense, the methods were accepted if the data confessed. The methods had to pass a series of tests. First, each method was examined to see if there were logical reasons to reject it. Next, the methods were tested on simulated data. Only if a method passed these two tests was it tried on real data. In evaluating the performance of methods on real data, two assumptions were made which should be repeated here to avoid misunderstanding: first, if two methods which are sensitive to different kinds of violations of assumptions give the same answer, it is fairly likely to be the right answer, and secondly, it is fairly unlikely that either Head Start programs or Title 1 programs could have a harmful effect.

Three methods failed before even being applied to data. To compare posttest scores of experimental and control groups is so obviously wrong that it was not even mentioned in the introduction, and is mentioned here only for
completeness. Such a comparison makes no attempt to control for initial differences between groups. The analysis of raw gain scores fails if there is differential growth rate in the two groups, and ordinary ANCOVA or multiple regression methods do not adequately correct for differential regression towards the mean in the groups.

A more competent statistician would have ruled out factor analysis regression at the first stage also, but since the technique had been advocated, and applied to one of the data sets I analyzed here (see Jurs, 1971), I pursued this technique to the point of trying it on a simulated data set. At this point it became clear that the method was inadequate to deal with the main problem of estimating treatment effects. It may give more stable estimates of regression coefficients than ordinary regression (at least the results are consistent with that conclusion) but it does not give unbiased estimates of treatment effects.

The analysis of standardized gain scores can give unbiased results in conditions found in many field studies, as can the method of comparing the covariate-pretest correlation with the covariate-posttest correlation. However, each of these methods is of fairly limited applicability since many studies used no pretest. Using ANCOVA corrected for unreliability will often prove to be the most useful technique, since any kind of covariate can
be used. Again, it should be kept in mind that the tabled \( F \) values are too liberal, and so a range of possible reliability estimates should be examined to see how stable the results are.

When correcting for unreliability, the DeGracie-Fuller method seems to give more conservative estimates than the Porter method (it seems to be less powerful), and in the studies analyzed here it would not have shown positive results even if they had occurred. This is not necessarily bad, but it should be kept in mind, especially when there are few subjects (and thus the test would lack power to an even greater extent). The Porter method seems to give nearly unbiased results when the right estimate of reliability is used. Internal consistency estimates are generally too high, while covariate-pcsttest correlations apparently give correct results. This conclusion is supported by several lines of evidence: the double pretest analysis of the Title 1 data, and the high prior probability that neither Head Start summer programs nor Title 1 programs had any effect, and especially not a negative effect. Furthermore, standardized gain score analysis and treatment-test correlation analysis of the Title 1 data support the hypothesis that the program was ineffective. Thus although the primary emphasis in this paper was on methodology, a tentative conclusion can be made that neither
Headstart nor Title 1 seemed to have any effect on the cognitive abilities tested.

One valuable lesson to be learned from this study is that the availability of many possible techniques for analyzing quasi-experiments is not a luxury, but a necessity. For several subsets of the data investigated here, one or more methods were unusable. Sometimes slopes were unequal so that ANCOVA could not be done, while other times simple factor models would not fit. There were floor effects which made gain score analysis inappropriate. Thus, researchers must sometimes try many methods in the hope that at least one of them might be informative to the researcher and to the researcher's audience.
Reference Notes

1. Boruch, R. F. Double pretests for checking certain threats to the validity of some conventional evaluation designs or stalking the null hypothesis. Unpublished manuscript, Northwestern University, 1976.


References


Campbell, D. T., & Boruch, R. F. Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C. A. Bennett, & A. Lumsdaine (Eds.), Central issues in social program evaluation. New York: Academic Press, in press.

Campbell, D. T., & Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth(Ed.), Compensatory education: A national debate,


Joreskog, K. G. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 1969,


Table 1. One-factor simulation.

<table>
<thead>
<tr>
<th></th>
<th>Post</th>
<th>Pre</th>
<th>Trt</th>
<th>Cov 1</th>
<th>Cov 2</th>
<th>Cov 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posttest</td>
<td>1.0000</td>
<td>.7465</td>
<td>.5492</td>
<td>.6658</td>
<td>.6198</td>
<td>.4984</td>
</tr>
<tr>
<td>Pretest</td>
<td>.8000</td>
<td>1.0000</td>
<td>.5435</td>
<td>.6282</td>
<td>.6378</td>
<td>.5055</td>
</tr>
<tr>
<td>Treatment</td>
<td>.5050</td>
<td>.5050</td>
<td>1.0000</td>
<td>.4674</td>
<td>.5296</td>
<td>.3395</td>
</tr>
<tr>
<td>Covariate 1</td>
<td>.6320</td>
<td>.6320</td>
<td>.3990</td>
<td>1.0000</td>
<td>.5083</td>
<td>.4348</td>
</tr>
<tr>
<td>Covariate 2</td>
<td>.6320</td>
<td>.6320</td>
<td>.3990</td>
<td>.5000</td>
<td>1.0000</td>
<td>.4401</td>
</tr>
<tr>
<td>Covariate 3</td>
<td>.6320</td>
<td>.6320</td>
<td>.3990</td>
<td>.5000</td>
<td>.5000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Hypothetical (population) values are below the diagonal, while sample values are above the diagonal.
Table 2. Regression of one-factor model.

<table>
<thead>
<tr>
<th></th>
<th>Beta Weights</th>
<th></th>
<th></th>
<th></th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cov 1</td>
<td>Cov 2</td>
<td>Ccv 3</td>
<td>Trt</td>
<td></td>
</tr>
<tr>
<td>Sample_Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L.S. Regression</td>
<td>.1909</td>
<td>.4220</td>
<td>.3043</td>
<td></td>
<td>.5744</td>
</tr>
<tr>
<td>L.S. Factor Analysis</td>
<td>.2422</td>
<td>.3597</td>
<td>.3199</td>
<td></td>
<td>.5708</td>
</tr>
<tr>
<td>Unbiased F.A.</td>
<td>.3214</td>
<td>.4773</td>
<td>.4245</td>
<td></td>
<td>.5708</td>
</tr>
<tr>
<td>L.S. Regression</td>
<td>.1791</td>
<td>.3805</td>
<td>.2626</td>
<td>.1565</td>
<td>.5926</td>
</tr>
<tr>
<td>L.S. Factor Analysis</td>
<td>.2104</td>
<td>.3202</td>
<td>.2895</td>
<td>.1638</td>
<td>.5898</td>
</tr>
<tr>
<td>Unbiased F.A.</td>
<td>.2693</td>
<td>.4098</td>
<td>.3705</td>
<td>.2696</td>
<td>.5898</td>
</tr>
<tr>
<td>Population_Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L.S. Regression</td>
<td>.3160</td>
<td>.3160</td>
<td>.3160</td>
<td></td>
<td>.5991</td>
</tr>
<tr>
<td>L.S. Factor Analysis</td>
<td>.3138</td>
<td>.3138</td>
<td>.3138</td>
<td></td>
<td>.5923</td>
</tr>
<tr>
<td>Unbiased F.A.</td>
<td>.4174</td>
<td>.4174</td>
<td>.4174</td>
<td></td>
<td>.5923</td>
</tr>
<tr>
<td>L.S. Regression</td>
<td>.2828</td>
<td>.2828</td>
<td>.2828</td>
<td>.1665</td>
<td>.6202</td>
</tr>
<tr>
<td>L.S. Factor Analysis</td>
<td>.2820</td>
<td>.2820</td>
<td>.2820</td>
<td>.1650</td>
<td>.6164</td>
</tr>
<tr>
<td>Unbiased F.A.</td>
<td>.3630</td>
<td>.3630</td>
<td>.3630</td>
<td>.2124</td>
<td>.6164</td>
</tr>
</tbody>
</table>
Table 3. Test of factor models on simulated data.

<table>
<thead>
<tr>
<th>Treatment size</th>
<th>One factor model</th>
<th>Two factor model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\chi^2$</td>
<td>$p$</td>
</tr>
<tr>
<td>Zero</td>
<td>5.73</td>
<td>.334</td>
</tr>
<tr>
<td>Small</td>
<td>4.13</td>
<td>.531</td>
</tr>
<tr>
<td>Medium</td>
<td>14.56</td>
<td>.012</td>
</tr>
<tr>
<td>Large</td>
<td>30.76</td>
<td>.000</td>
</tr>
</tbody>
</table>

Note: The factor loadings listed are for the second factor only. The degrees of freedom for the chi-squares are 3 and 5 for the two and one factor models respectively.
Table 4. Means and variances of real data.

<table>
<thead>
<tr>
<th>Title 1</th>
<th>Metropolitan</th>
<th></th>
<th>Experimental</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Posttest</td>
<td>Pretest</td>
<td>Posttest</td>
<td>Pretest</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>Variance</td>
<td>Mean</td>
<td>Variance</td>
</tr>
<tr>
<td>SES scale</td>
<td>.000</td>
<td>.351</td>
<td>-.006</td>
<td>.298</td>
</tr>
<tr>
<td>Stanford</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posttest</td>
<td>29.17</td>
<td>85.02</td>
<td>23.50</td>
<td>37.34</td>
</tr>
<tr>
<td>Pretest</td>
<td>19.17</td>
<td>29.14</td>
<td>23.50</td>
<td>16.25</td>
</tr>
<tr>
<td>SES scale</td>
<td>.023</td>
<td>.388</td>
<td>-.302</td>
<td>.369</td>
</tr>
<tr>
<td>Headstart</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whites</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITPA</td>
<td>19.97</td>
<td>13.85</td>
<td>19.26</td>
<td>13.54</td>
</tr>
<tr>
<td>MRT</td>
<td>9.83</td>
<td>6.96</td>
<td>9.30</td>
<td>6.82</td>
</tr>
<tr>
<td>SES scale</td>
<td>.149</td>
<td>.527</td>
<td>-.154</td>
<td>.455</td>
</tr>
<tr>
<td>Blacks</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITPA</td>
<td>17.26</td>
<td>8.94</td>
<td>18.79</td>
<td>9.47</td>
</tr>
<tr>
<td>MRT</td>
<td>7.70</td>
<td>5.85</td>
<td>8.37</td>
<td>6.10</td>
</tr>
<tr>
<td>SES scale</td>
<td>.029</td>
<td>.547</td>
<td>-.024</td>
<td>.474</td>
</tr>
</tbody>
</table>
Table 5. Title 1 Metropolitan, pretest as covariate.

<table>
<thead>
<tr>
<th>Reliability</th>
<th>Porter F</th>
<th>$\Delta Y - \beta \Delta X / R$</th>
<th>DeGracie-Fuller F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>13.69**</td>
<td>-4.14</td>
<td>13.67**</td>
</tr>
<tr>
<td>.9</td>
<td>11.34**</td>
<td>-3.78</td>
<td>11.24**</td>
</tr>
<tr>
<td>.8</td>
<td>8.72**</td>
<td>-3.32</td>
<td>8.36**</td>
</tr>
<tr>
<td>.7</td>
<td>5.86*</td>
<td>-2.74</td>
<td>5.20*</td>
</tr>
<tr>
<td>.6</td>
<td>2.94</td>
<td>-1.95</td>
<td>2.26</td>
</tr>
<tr>
<td>.5</td>
<td>.52</td>
<td>-.86</td>
<td>.33</td>
</tr>
</tbody>
</table>

* Denotes significance at the .05 level.

** Denotes significance at the .01 level.
Table 6. Title 1 Stanford, pretest as covariate.

<table>
<thead>
<tr>
<th>Reliability</th>
<th>Porter F</th>
<th>$\frac{\Delta Y - \bar{Y} \Delta \bar{X}}{R}$</th>
<th>DeGracie-Fuller F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>4.19*</td>
<td>-1.74</td>
<td>4.18*</td>
</tr>
<tr>
<td>.9</td>
<td>2.35</td>
<td>-1.31</td>
<td>2.32</td>
</tr>
<tr>
<td>.8</td>
<td>.79</td>
<td>-0.76</td>
<td>.75</td>
</tr>
<tr>
<td>.7</td>
<td>.00</td>
<td>-0.06</td>
<td>.00</td>
</tr>
<tr>
<td>.6</td>
<td>.98</td>
<td>.87</td>
<td>.61</td>
</tr>
<tr>
<td>.5</td>
<td>6.03*</td>
<td>2.18</td>
<td>2.56</td>
</tr>
</tbody>
</table>

* Denotes significance at the .05 level.

** Denotes significance at the .01 level.
Table 7. Title 1 Stanford, SES scale as covariate.

<table>
<thead>
<tr>
<th>Reliability</th>
<th>Porter $F$</th>
<th>$\bar{Y} - \bar{X}/R$</th>
<th>DeGracie-Fuller $F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>12.02**</td>
<td>-3.83</td>
<td>12.00**</td>
</tr>
<tr>
<td>.9</td>
<td>10.74**</td>
<td>-3.62</td>
<td>10.70**</td>
</tr>
<tr>
<td>.8</td>
<td>9.24**</td>
<td>-3.37</td>
<td>9.13**</td>
</tr>
<tr>
<td>.7</td>
<td>7.49**</td>
<td>-3.04</td>
<td>7.24**</td>
</tr>
<tr>
<td>.6</td>
<td>5.46*</td>
<td>-2.60</td>
<td>5.05*</td>
</tr>
<tr>
<td>.5</td>
<td>3.20</td>
<td>-1.99</td>
<td>2.71</td>
</tr>
<tr>
<td>.4</td>
<td>1.03</td>
<td>-1.07</td>
<td>1.76</td>
</tr>
<tr>
<td>.3</td>
<td>.41</td>
<td>.47</td>
<td>.15</td>
</tr>
<tr>
<td>.2</td>
<td>7.54**</td>
<td>3.54</td>
<td>.90</td>
</tr>
</tbody>
</table>

* Denotes significance at the .05 level.

** Denotes significance at the .01 level.
Table 8. Title 1 Stanford, double pretest.

<table>
<thead>
<tr>
<th>Reliability</th>
<th>Porter F</th>
<th>$\Delta X - \bar{\Delta X}/5$</th>
<th>DeGracie-Fuller F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>12.70**</td>
<td>-2.39</td>
<td>12.68**</td>
</tr>
<tr>
<td>.9</td>
<td>11.63**</td>
<td>-2.29</td>
<td>11.61**</td>
</tr>
<tr>
<td>.8</td>
<td>10.37**</td>
<td>-2.17</td>
<td>10.29**</td>
</tr>
<tr>
<td>.7</td>
<td>8.86**</td>
<td>-2.01</td>
<td>8.67**</td>
</tr>
<tr>
<td>.6</td>
<td>7.03**</td>
<td>-1.80</td>
<td>6.69*</td>
</tr>
<tr>
<td>.5</td>
<td>4.85*</td>
<td>-1.50</td>
<td>4.35*</td>
</tr>
<tr>
<td>.4</td>
<td>2.40</td>
<td>-1.07</td>
<td>1.93</td>
</tr>
<tr>
<td>.3</td>
<td>.34</td>
<td>-.33</td>
<td>.26</td>
</tr>
<tr>
<td>.2</td>
<td>2.22</td>
<td>1.14</td>
<td>.33</td>
</tr>
</tbody>
</table>

* Denotes significance at the .05 level.

** Denotes significance at the .01 level.
Figure 1. Scattergram of posttest versus pretest.
Figure 2. Effects of adding error to covariate.
Acknowledgments

Thanks to Dr. Robert Boruch and his grant from the National Institute of Education (NIE-C-74-0115) for providing intellectual, moral, and monetary support for this dissertation.

I would also like to thank Dr. Donald I. Campbell, whose writings provided enlightenment and whose grant from NSF (SOC71-03704) provided the funds to obtain the Head Start data.

My doctoral committee (Dr. Wayne Bartz, Dr. Leroy Wolins, Dr. Fred Brown, Dr. Ron Peters, and Dr. Ellis Hicks) has provided guidance and assistance throughout my years of graduate training.