This is to certify that the

thesis entitled

RESIDUAL GAIN SCORES AS A
CRITERION FOR CHANGE:
INFERENTIAL PROBLEMS

presented by

Khalil Elaian

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Educational Statistics and Research Design; Department of Counseling, Educational Psychology and Special Education

Major professor

Date 1/23/84

RESIDUAL GAIN SCORES AS
A CRITERION FOR CHANGE:
INFERENTIAL PROBLEMS

by

Khalil Elaian

A DISSERTATION

submitted to
Michigan State University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology,
and Special Education

1984

ABSTRACT

RESIDUAL GAIN SCORES AS A CRITERION FOR CHANGE: INFERENTIAL PROBLEMS

by

Khalil Elaian

To determine the effect of teacher behaviors, W, in process-product research, residual gain scores, Z, are often used as the criterion. Significant correlations between class means of residualized gain scores and teacher behaviors, $r_{\bar{Z}W}$'s, are taken as evidence of teacher effects. The purposes of the study were to determine the conditions under which testing H_0: $\rho_{\bar{Z}W} = 0$ is, in fact, equivalent to testing for no teacher behavior effect, and also to investigate the appropriateness of using different definitions of residual gain scores in testing the null hypothesis. Five different forms of residualized gains were considered based on the total, Z_1, between, Z_2, and within regression coefficient, Z_3, a newly derived estimate of the regression of posttest class effects on pretest class effects, Z_4, and finally the parameter for the class effects regression coefficient, Z_5.

A linear structural model was built to determine the conditions under which testing $\rho_{\bar{Z}W} = 0$ is equivalent to testing no teacher behavior effect on student achievement. The analytic results showed that the two null hypotheses are equivalent if either of the following conditions are met: (a) there is no initial confounding of teacher behavior and class composition or (b) the slope of posttest class effect on pretest class effect, β_1, is equal to the slope of

posttest on pretest for within classes, β_2 and a perfectly reliable pretest. When the conditions are not met; however, the two null hypotheses are equivalent only for Z_4 and Z_5.

A Monte Carlo approach was taken to investigate the appropriateness of using different $r_{\bar{z}w}$'s in testing the hypothesis of no teacher behavior effect. Three criteria were considered: (a) the mean estimates of $\rho_{\bar{z}w}$'s, (b) empirical Type I error rates, aand (c) empirical power. Parameters varied in the study were the degree of initial confounding, the reliability of the pretest, the number of classrooms, and the number of students in a classroom.

The results of the study showed that when there was a substantial amount of initial confounding, the test statistics using $r_{\bar{z}_1w}$, $r_{\bar{z}_2w}$, and $r_{\bar{z}_3w}$ were only valid in a few situations. These tests, particularly the tests using $r_{\bar{z}_1w}$ and $r_{\bar{z}_3w}$, tended to be too liberal in situations where $\beta_1 = \beta_2$ or $\beta_1 > \beta_2$ and too conservative when $\beta_1 < \beta_2$. Parallel results for the tests using $r_{\dot{z}_1w}$ and $r_{\dot{z}_3w}$ were obtained with increasing sample size. However, the test statistics using $r_{\dot{z}_4w}$ and $r_{\dot{z}_5w}$ were the only tests which remained valid as initial confounding, sample, and class size increased and in the presence of errors of measurement. Also, the results indicated that increasing sample and class size increased the empirical power of both $r_{\bar{z}_4w}$ and $r_{\bar{z}_5w}$ in situations where $\beta_1 = \beta_2$ or $\beta_1 > \beta_2$.

It was concluded that procedures used by process-product researchers in forming residual gain scores typically provide misleading results. Sometimes the test statistics used are too liberal and other times they are too conservative. Therefore, it is recommended that process-product researchers who wish to test for no teacher behavior effect use Z_4. In addition to yielding valid Type I error rates across all conditions investigated, the procedure had reasonable power and does not have the unrealistic requirement of knowing the value of a parameter a priori.

DEDICATION

To my wife,

Nasrin Bakir,

and my son,

Rami Elaian.

ACKNOWLEDGEMENTS

I am deeply grateful to Professor Andrew C. Porter, my academic advisor and dissertation chairperson, for his generous and endless advice, encouragement, support, and patience. I would like to thank Professor Richard Houang for serving as a member of my dissertation committee and for the generous help he has given me with the dissertation. I would also like to thank Professors Robert Floden and James Stapleton for serving on my committee.

Working in the Office of Research Consultation (ORC) has provided me with invaluable experience. Many thanks to Professor Joe L. Byers who hired me in the ORC.

I would further like to thank the University of Jordan for four years of financial support.

I wish to acknowledge the moral support of my parents, wife, and friends.

Finally, I wish to thank Barbara Reeves for her typing of the final manuscript.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

STATEMENT OF THE PROBLEM

Residual gain scores (RGS) are often used as a criterion for measuring change in educational research. For example, in reference to evaluating teacher effectiveness, Veldman and Brophy (1974) state "it is generally accepted that residual gain scores are superior to simple pretest-posttest difference scores as measures of teacher influence" (p. 321). Process/product research on teaching can be used to illustrate the practice of setting residual gain scores as a criterion for study (Gage, 1977). Process refers to teaching behavior and product refers to student learning. Residual gain scores are used as a product variable which is meant to control for initial differences among classrooms in their compositions of students. The residual gain scores are typically constructed from student pre- and post-instruction achievement test scores.

Brophy and Evertson's (1974) two year replicated study conducted at the University of Texas provides a specific example of using residualized gain scores as the criterion in process/product research. Thirty teachers were included in the first year, and 28 in the second year. Classroom observations were made to assess teacher behavior. Scores on five subtests of the Metropolitan Achievement Test (MAT) were available for each student. The MAT obtained on the first year was used as the pretest and the MAT for the second year as posttest. For each student, predicted values of the posttest scores were determined from the pretest scores based on the total sample regression line. Residual gain scores were computed by subtracting predicted posttest scores from the actual posttest scores.

To determine the effects of teacher behaviors, Pearson's product-moment correlation coefficients between process variables and average residual gain scores, aggregated by teacher, were obtained. The sample correlations were tested for significance by t-tests, with c - 2 degrees of freedom, where c refers to the number of teachers. The null hypothesis that Brophy and Evertson intended to test was that teacher behavior had no effect on student achievement. The aim of the present study is to investigate the appropriateness of using residualized gain scores in order to determine the process product relationship.

<u>Definition of Residual Gain Scores</u>

Consider the model used in forming residual gain scores:

(1) $Z = Y(t) - K\,Y(0)$ where $Y(0)$ is the measure at time 0,
$Y(t)$ is the measure at time t, and
K is an adjustment coefficient.

As described previously, $\bar{Z}$, the constructed residual gain score aggregated by teacher, is correlated with a measure of teacher behavior, W. Let $r_{\bar{Z}W}$ denote the sample correlation coefficient between $\bar{Z}$ and W, and $\rho_{\bar{Z}W}$ be the corresponding parameter. For a test of H_0: $\rho_{\bar{Z}W} = 0$ to be appropriate, not only must the variables, Z and W, be linearly related but, in addition, $\rho_{\bar{Z}W}$ must be only a function of change in achievement caused by W. If either one of the above conditions is false, a test of H_0: $\rho_{\bar{Z}W} = 0$ will lead to spurious conclusions.

As Z is constructed from equation 1, the appropriateness of $\rho_{\bar{Z}W} = 0$ for a null hypothesis depends upon the choice of K, the adjustment coefficient. While K is assumed to be a known constant, in practice, this is seldom the case. Usually K is estimated from the relationship between Y (0) and Y (t), in terms of a regression coefficient. Because the nature of the data on student performance is hierarchical (i.e., students are nested within classrooms) three regression coefficients are available: the between classroom regression coefficient, the within classrooms regression coefficient, and the total regression coefficient. For most educational data, these three regression coefficients are not interchangeable. Further, it will be shown that in some situations, none of the three coefficients estimate an appropriate correction parameter.

To further complicate matters, the sampling distribution of $r_{\bar{z}w}$ will be a function of the estimator for K. Unfortunately, the nature of the sampling distribution of $r_{\bar{z}w}$ is unknown (at least for most situations), and the use of the t-distribution to test H_o : $\rho_{\bar{z}w} = 0$ as in Brophy and Evertson's study, may not be valid even when the sample regression coefficient estimates an appropriate correction parameter (Draper & Smith, 1981).

Research Questions

The intent of the present study was to investigate the appropriateness of using a t-test to test $\rho_{\bar{z}w} = 0$ as evidence for no teacher behavior effect on student achievement. More specifically, the following research questions were investigated.

1. What are the conditions under which testing $\rho_{\bar{z}w} = 0$ is equivalent to testing no teacher behavior effect for different methods of defining Z?
2. Given no teacher behavior effect, how well does the distribution of the "t" statistic based on each of several different methods of defining z approximate the theoretical t-distribution for varying amounts of (a) initial confounding, (b) presence of errors of measurement in the premeasure, (c) number of classrooms, and (d) class size?

The investigation was conducted in two steps. First, the conditions under which $\rho_{\bar{z}w} = 0$, if and only if there is no effect of teacher behavior on student achievement, were determined analytically. Second, a simulation study was conducted to investigate empirically the distribution of "t" statistics using different methods of testing $r_{\bar{z}w}$.

In Chapter II, relevant literature will be reviewed. In order to examine the situation thoroughly, a structual model is introduced in Chapter III. Chapter IV presents the design of the simulation study. The results obtained from the empirical study and the conclusions reached are presented in Chapters V and VI.

CHAPTER II

REVIEW OF THE LITERATURE

In experimental research the experimenter manipulates variables of interest and observes the manner in which the manipulation affects the variation of the dependent variables. In order to be reasonably sure that the observed variation in the dependent variable is indeed due to the manipulated variables, the experimenter must control all possible confounding variables. Porter & Chibucos (1974) suggested two catergories for these possible confounding variables in the context of program evaluation:

> 1. Systematic differences in the dependent variable dimensions that are present in the units of analysis at the outset of program participation.
>
> 2. Systematic differences that occur in the dependent variable dimensions during program participation which are not a function of program participation. (p. 440).

While randomization is one of the most powerful methods to control confounding variables of the first category it does not insure controlling confounding variables of the second category. To the extent both categories of possible confounding variables are controlled, arguments for causal relationships between independent and dependent variables are strengthened.

Studies of natural variation are also used to investigate the possibility of causal relationships among independent and dependent variables. As was the case for experimental research the investigator must be concerned about both types of confounding variables. In studies of natural variation, however, randomization is by definition not a part of the design and so other methods must be employed to guard against confounding. One general method, which has enjoyed considerable use, involves the formulation of an index of response of which residualized gain scores, the focus of this study, represented a specific type.

Alternative Indices of Responses

The index of response is defined by $Z_{ij} = y(t)_{ij} - K\ y(0)_{ij}$ where Y(0), Y(t) are pre and post measures for the i^{th} individual in the j^{th} group, K is some known constant. In addition to requiring scores on the measure of interest at two points in time to formulate Z, K has to be set to an apriori known value. However, the value K should take depends on knowledge regarding the natural growth model which adequately describes the data if there were no effects of the independent variable.

The most commonly used index of response is raw gain scores, where K is set to unity,

$D_{ij} = y(t)_{ij} - y(0)_{ij}$ where D_{ij} is the raw gain for the i^{th} individual in the j^{th} group.

In other words, raw gain scores are created by taking the difference of the post measure and premeasure scores on the dependent variable dimension.

Raw gain scores as a measure of individual change have been criticized in the literature for having low reliability and for correlating negatively with premeasure scores (Cronbach & Furby, 1970, Linn & Slinde, 1977; Lord, 1963). Cronbach and Furby, have also questioned the use of raw gain as a strategy to measure group change in studies of natural variation, agreeing with Lord (1967) that gain scores are an inappropriate strategy to control for confounding variables in natural variation studies of causal relationships. In contrast, Porter (1973) has suggested that under certain assumptions gain scores may provide the best technique for natural variation studies. Porter argued that given treatment effects are additive the pre and posttest measure the same variable in a common metric and there is no change in variances from pretest to posttest; it can be shown that the gain score strategy does provide a reasonable approach to data analysis in natual variation studies. Bryk and Weisberg (1977) showed that under natural growth (i.e., no treatment effect) this gain score strategy provides an unbiased estimate of the treatment if and only if the group growth patterns are parallel (which is equivalent to Porter's assumptions).

Standardized gain scores represent yet another form of index of response that has been used to analyze data from studies of natural variation. K in the index of response is set to either one of $\sigma_{y_t} / \sigma_{y_o}$, $\sigma_{T_{y_t}} / \sigma_{T_{y_o}}$ or s_{y_t} / s_{y_o} where $\sigma^2_{y_t}$ and $\sigma^2_{y_o}$ are the population variances of the dependent variable dimension at pre and posttests, $\sigma^2_{T_{y_t}}$ and $\sigma^2_{T_{y_o}}$ are the population variances for the true variables and s_{y_t} and s_{y_o} are the sample estimates of σ_{y_t} and σ_{y_o}. Using ANOVA of standardized gain scores as a strategy to control initial confounding was introduced by Kenny (1975). Even though Kenny did not distinguish between the different types of standardized gain scores, he argued that when individuals were assigned to a program based on sociological or demographic variables, standardized gain scores provide the best index of response for controlling initial confounding. Olejnik and Porter (1981) clarified Kenny's recommendations by showing that the validity of standardized gain scores is dependent upon the model of natural growth that applies in the absence of treatment effects. They also pointed out that the two alternative ratios of population standard deviations are equivalent if the reliability of the pretest and posttest are equal. Finally and perhaps most importantly, they pointed out that using a ratio of sample standard deviations followed by ANOVA is an incorrect procedure that results in misleadingly small standard errors.

Residual gain scores are yet another form of index of response that has been used in studies of natural variation. Three different types of residual gain scores appear in the literature of measuring change. The first, which is called True residual gain scores is defined by setting K in the index of response to the slope of true posttest on the true pretest. True residual gain scores were suggested by Tucker et al. (1966) and called a "base free measure of change." The second, called observed residual gain scores, Z, sets K = $\beta_{y_t y_o}$ (i.e. the slope of the manifest variables). The third called estimated residual gain scores, Z, sets K=$\beta_{y_t y_o}$ where $\beta_{y_t y_o}$ is the sample estimate of the slope of y_t on y_o.

Residual gain scores as measures of individual change have been characterized in the literature as uncorrelated with initial status but suffering from low reliability (Kessler, 1977; Linn & Slinde, 1977). Using ANOVA on observed residual gain scores as an analysis strategy in natural variation studies is comparable to using analysis of covariance. The only difference between the two procedures is that ANCOVA estimates the value of K from the data while ANOVA on the observed residual gain scores requires that K be set apriori to $\beta_{y_t y_o}$.

ANOVA on true residual gain scores is parallel to estimated true scores analysis of covariance originally developed by Porter (1967). Again the distinction is that the true residual gain score approach requires that a population slope be known apriori while estimated true score ANOVA estimates that slope from the data. Performing ANOVA on the estimated residual gain scores raises at least two problems. First the expected value of the estimated residual gain score is unknown (Draper and Smith, 1981) making it difficult to determine whether the strategy provides unbiased estimates of the causal relationship of interest. Second, the procedure suffers from the same bias of standard errors that Olejnik and Porter (1981) noted for standardized gain scores using sample standard deviations.

Uses of Indices of Responses

Using an index of response in lieu of randomization in natural variation studies has been a controversial topic. Perhaps the best known antagonist of their use is Lord (1967, 1969) who has stated "with the data usually available for such studies, there is simply no logical or statistical procedure that can be counted on to make proper allowance for uncontrolled preexisting differences between groups" (Lord, 1967, p.35). More recently Cronbach and Furby (1970) have indicated basic agreement with Lord's pessimistic view of the utility of using statistical adjustment in natural variation studies. On the other hand, Elashoff, 1969; Hornquist, 1968; Porter & Chibucos, 1974 hold a more optimistic view. Hornquist (1968) has stated

> Even if the initial standing of the subjects is controlled by means of a number of relevant variables, there will always be room for uncontrolled differences that may be important. The investigator, who because of the nature of his problem cannot use random or systematic assignments of subjects to treatments, has to live with an insecurity in that respect . . . and try to behave intelligently within the limitations of his design . . . or leave the scene of non-experimental research"(p.57).

Porter (1973) has stated ". . . the interpretation of results from designs lacking random assignment requires yet another degree of tentativeness above and beyond what would have been required had random assignment been employed" (p.41).

Research on teacher effectiveness is one of the areas in which residual gain scores have been used most heavily. For some researchers (e.g. Rosenshine,1970) residual gain scores are considered the definition for teacher effectiveness and so the logical dependent variable in studies to identify desirable teacher behaviors. Known as process/product research (Dunkin & Biddle, 1974), studies of effective teacher behavior obtain pre and posttests of students achievement to form the dependent variable and observations of teachers to form independent variables. The residual gain scores are computed for each student and then aggregated to the classroom/teacher level. The correlations of class means on residual gain scores and teacher behaviors are computed and tested for significance. Significant correlations are taken as evidence that teacher behavior affects student achievement. Examples of process/product research using residualized gain scores to control for confounding variables are Brophy & Evertson, 1974; Creemer, 1974; Creemer and Weeda, 1974; Soar, 1966; and Veldman and Brophy, 1974. In all of these studies $\beta_{y_t y_o}$ was unknown and so estimated to define the "constant" in the residualized gain scores. The researchers, however, ignored this distinction when conducting their tests of significance of correlation between teacher behavior and residualized gains. A test statistic using r_{zw}, which is appropriate to test H_o : $\rho_{\bar{z}w} = 0$ does not necessarily imply that the parallel test statistics using $r_{\bar{z}w}$, is also a valid test of $\rho_{\bar{z}w} = 0$.

Testing $\rho_{\bar{z}w}=0$ as a test for no teacher behavior effect was investigated in the present study. The investigation was in two parts, analytic and empirical. The analytic part was conducted to determine the conditions under which $\rho_{\bar{z}w}=0$ is equivalent to testing no effect of a teacher behavior on student achievement. The investigation considered several different possible formulations of Z. The empirical investigation was conducted to investigate the appropriateness of a "t" test statistic to test H_o : $\rho_{\bar{z}w}=0$ when sample estimates rather than population parameters were used to define the residualized gain scores. A Monte Carlo method was used to simulate the sampling distributions of the different test statistics based on different formulations of residualized gain scores. These

were then compared to the theoretical reference distributions to determine the validity of each test statistic under study.

CHAPTER III

THE ANALYTIC CHAPTER

In this chapter, a linear structural model that defines the problem of measuring change in studies of process/product research will be presented. The model incorporates the aggregated characteristics of the data and the possibility of measurement errors. Given the model, the conditions under which $\rho_{\bar{Z}W} = 0$ is equivalent to no teacher behavior effect on achievement will be identified.

A Linear Structural Model for Process/Product Research

As in equation 1, residual gain scores are constructed from Y(0) and Y(t), the pre- and post-measures of student achievement. The proposed structural model attempts to elucidate the relationships among Y(0), Y(t), and W, a variable representing teacher behavior.

For student i in class j, the observed score $Y(L)_{ij}$ can be decomposed into:

(2) $Y(L)_{ij} = \eta(L)_{ij} + e(L)_{ij}$, L = 0, t

where $\eta(L)_{ij}$ is the part of $Y(L)_{ij}$ which is free from errors of measurement, and $e(L)_{ij}$ represents measurement error. The $\eta(L)_{ij}$ is further decomposed into two components: the class effect and the deviation of student score from his class mean,

(3) $\eta(L)_{ij} = A(L)_j + V(L)_{ij}$, L = 0, t

where $A(L)_j$ is the class effect at time L, and $V(L)_{ij}$ represents the deviation of the i^{th} student score from the mean of j^{th} class. Combining the two equations, $Y(L)_{ij}$ can be written as

(4) $Y(L)_{ij} = A(L)_j + V(L)_{ij} + e(L)_{ij}$, L = 0 ,t.

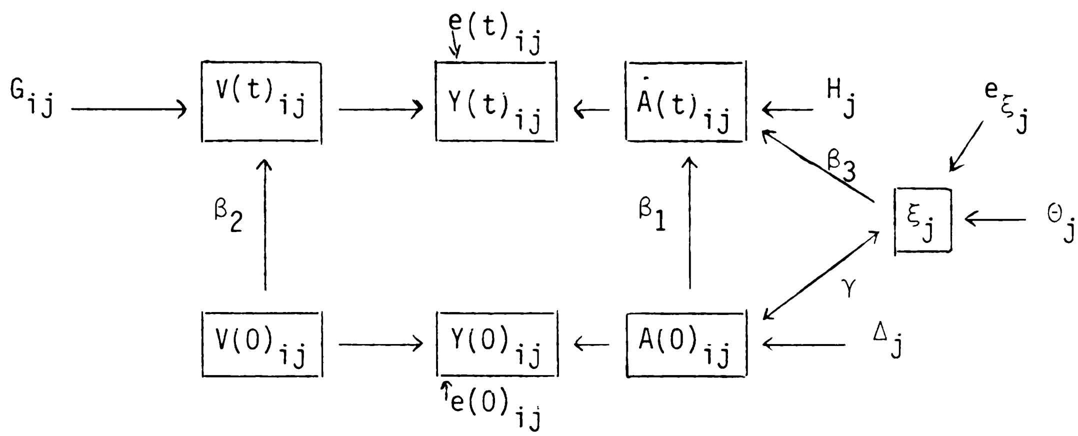
The measure of teacher behavior can also be decomposed into

(5) $W_j = j + e_{\xi_j}$,

where ξ_j is the true measure of the behavior of teacher j assigned to class j and e_{ξ_j} represents measurement error.

Schematically, the structural relationships among the three variables are shown in Figure 1.

Figure 1. A structural model.



The β's are the structural coefficients, γ represents the reciprocal relationship between ξ_j and $A(0)_j$. H_j, G_{ij}, Θ_j and Δ_j are residuals or specification errors. The structural equations for $A(t)_j$ and $V(t)_{ij}$ are

(6) $A(t)_j = \beta_1 A(0)_j + \beta_3 \xi_j + H_j$.

(7) $V(t)_{ij} = \beta_2\, V(0)_{ij} + G_{ij}$.

Within class j, $V(t)_{ij}$ is linearly related to $V(0)_{ij}$. This is equivalent to the assumption of a linear growth operating within each class at the individual level. The same rate of growth, β_2, occurs within each class.

The decomposition of $\eta(L)_{ij}$ into $A(L)_j$ and $V(L)_{ij}$ also implies that the class effect is additive (i.e., $A(L)_j$ is a constant effect for all students in the same class). The effect of the teacher behavior, W, on student achievement is the

same for all students in the same class. Teacher behavior may, however, have a direct effect (β_3) on A(t) and a reciprocal relationship (γ) with A(0). The former will result in changes in performance (for the class as a whole) as a consequence of being exposed to the teacher behavior of interest. The reciprocal relationships (γ) represents confounding between initial class composition and teacher behavior. In school settings, students are virtually never randomly assigned to classes, and so substantial class effects exist before the start of the school year. Importantly, these differences may be at least in part a consequence of having teacher j in class j. This will have some impact on $A(t)_j$ through A(0). Also, this reciprocal relationship represents the possibility that the composition of the class may affect the way the teacher teaches (Doyle, 1979) which can affect $A(t)_j$.

Given the following two assumptions, β_3 represents the effect of the teacher behavior on student achievement:

1. Prior to the study, there is no other teacher behavior, ξ_1, that is correlated with ξ and which has some effect on A(0)j and/or A(t)j.
2. During the study, there is no other teacher behavior, ξ_2, that is correlated with ξ and which has some effect on A(t)j.

These first two assumptions are necessary to leave the interpretation of $\beta_3 \neq 0$ clearly a function of the effect of W and not some other teacher behavior variables.

The Relationship Between $\rho_{\hat{z}w}$ and β_3

The observed variables Y_t, Y_o and W are assumed to have a multivariate normal distribution with a mean vector of zero and a variance covariance matrix, Σ (see Table 1).

Table 1
The Total Variance - Covariance Matrix (Σ)

	Y(t)	Y(0)	W
Y(t)	$\sigma^2_{A_t} + \sigma^2_{V_t} + \sigma^2_{e_t}$		
Y(o)	$\beta_1\sigma^2_{A_0} + \beta_3\gamma\sigma_{A_0}\sigma_\xi + \beta_2\sigma^2_{V_0}$	$\sigma^2_{A_0} + \sigma^2_{V_0} + \sigma^2_{e_0}$	
W	$\beta_1\gamma\sigma_{A_0}\sigma_\xi + \beta_3\sigma^2_\xi$	$\gamma\sigma_{A_0}\sigma_\xi$	$\sigma^2_\xi + \sigma^2_{e_\xi}$

In the structural model, errors of measurement and specification errors are assumed to be uncorrelated among themselves and with the latent variables, 's, V's, A's and ξ.

The coefficient, ρ_{zw} can be written as

$$\rho_{\bar{z}w} = \frac{\sigma_{\bar{z}w}}{\sigma_{\bar{z}}\sigma_w}$$

To determine the relationship between ρ_{zw} and β_3, the variances and covariance are expressed in terms of the structural coefficients.

The covariance between Z and W can be written as

$$\begin{aligned}\sigma_{\bar{z}w} &= E(\bar{z}w) - E(\bar{z})E(w)\\ &= E(\bar{y}_{(t)}w) - KE(\bar{y}_{(0)}w)\\ &= E(A_t + \bar{V}_t + \bar{e}_t)(\xi + e_\xi) - KE(A_0 + \bar{V}_0 + \bar{e}_0)(\xi + e_\xi)\\ &= E(A_t\xi) + E(\bar{V}_t\xi) - KE(A_0\xi) - KE(\bar{V}_0\xi)\end{aligned}$$

Since V's are defined at the individual level and ξ at the class level,

$$E(V(L)\xi) = EE_j(\xi \sum_{i=1}^{s} V(L)_{ij})$$

$$= E(\xi_j \sum_{i=1}^{s} E_j \frac{(V(L)_{ij})}{s}) = E_j(\xi_j(0)) = 0$$

Thus,

Table 2

Relations of the Different Regression Coefficients and Coefficients and $\rho_{\dot{z}w}$ to the Structural Coefficients

Regression Coefficient	Notation	Definitions	Relations to the Structural Coefficients	$\rho_{\dot{z}w}$ when $\beta_3 = 0$	$\rho_{\dot{z}w}$ when $\beta_3 \neq 0$
Total Regression Coefficient	β_T	$\frac{\Sigma\Sigma_j(Y(t)_{ij} - M_y(t))(Y(0)_{ij} - M_{y(0)})}{\Sigma(Y(0)_{ij} - M_{y(0)})^2}$	$\frac{\beta_1\sigma^2_{A_0} + \beta_2\sigma^2_{V_0} + \gamma\beta_3\sigma_{A_0}\sigma_\xi}{\sigma^2_{A_0} + \sigma^2_{V_0} + \sigma^2_{e_0}}$	$\frac{\sigma_{A_0}\sigma_\xi}{\sqrt{\sigma^2_{\dot{z}}\sigma^2_w}}\left(\frac{(\beta_1 - \beta_2)\sigma^2_{V_0} + \beta_1\sigma^2_{e_0}}{\sigma^2_{A_0} + \sigma^2_{V_0} + \sigma^2_{e_0}}\right)$	$\frac{1}{\sqrt{\sigma^2_{\dot{z}}\sigma^2_w}}\left(\frac{\gamma\sigma_{A_0}\sigma_\xi((\beta_1 - \beta_2)\sigma^2_{V_0} + \beta_1\sigma^2_e - \beta_3\gamma\sigma^2_{A_0}\sigma_\xi)}{\sigma^2_{A_0} + \sigma^2_{V_0} + \sigma^2_{e_0}} + \beta_3\sigma^2_\zeta\right)$
Between Regression Coefficient	β_B	$\frac{\Sigma(\bar{Y}(t)_j - M_y(t))(\bar{Y}(0)_j - M_{y(0)})}{\Sigma(\bar{Y}(0)_j - M_{y(0)})^2}$	$\frac{\beta_1\sigma^2_{A_0} + \frac{1}{s}\beta_2\sigma^2_{V_0} + \gamma\beta_3\sigma_{A_0}\sigma_\xi}{\sigma^2_{A_0} + \frac{1}{s}\sigma^2_{V_0} + \frac{1}{s}\sigma^2_{e_0}}$	$\frac{\gamma\sigma_{A_0}\sigma_\xi}{\sqrt{\sigma^2_{\dot{z}}\sigma^2_w}}\left(\frac{(\beta_1 - \beta_2)\sigma^2_{V_0} + \beta_1\sigma^2_{e_0}}{s\sigma^2_{A_0} + \sigma^2_{V_0} + \sigma^2_{e_0}}\right)$	$\frac{1}{\sqrt{\sigma^2_{\dot{z}}\sigma^2_w}}\left(\frac{\gamma\sigma_{A_0}\sigma_\xi((\beta_1 - \beta_2)\sigma^2_{V_0} + \beta_1\sigma^2_{e_0} - s\beta_3\gamma\sigma_{A_0}\sigma_\xi)}{s\sigma^2_{A_0} + \sigma^4_{V_0} + \sigma^2_e} + \beta_3\sigma^2_\xi\right)$
Within Regression Coefficient	β_W	$\frac{\Sigma_j(Y(t)_{ij} - M_y(t)_j)(Y(0)_{ij} - M_{y(0)_j})}{\Sigma_j(Y(0)_{ij} - M_{y(0)_j})^2}$	$\beta_2\left(\frac{\sigma^2_{V_0}}{\sigma^2_{V_0} + \sigma^2_{e_0}}\right)$	$\frac{\gamma\sigma_{A_0}\sigma_\xi}{\sqrt{\sigma^2_{\dot{z}}\sigma^2_w}}\left(\frac{(\beta_1 - \beta_2)\sigma^2_{V_0} + \beta_1\sigma^2_{e_0}}{\sigma^2_{V_0} + \sigma^2_{e_0}}\right)$	$\frac{1}{\sqrt{\sigma^2_{\dot{z}}\sigma^2_w}}\left(\frac{\gamma\sigma_{A_0}\sigma_\zeta((\beta_1 - \beta_2)\sigma^2_{V_0} + \beta_1\sigma^2_{e_0})}{\sigma^2_{V_0} + \sigma^2_{e_0}} + \beta_3\sigma^2_\xi\right)$
	β_1	$\frac{\sigma A(t)A(0)}{\sigma^2_{A_0}}$	$\beta_1 = \frac{\beta_T\sigma^2_{Y_0} - \beta_2\sigma^2_{V_0}}{\sigma^2_{A_0}}$	0	$\frac{1}{\sqrt{\sigma^2_{\dot{z}}\sigma^2_w}}\left(\frac{\beta_1\sigma^2_{A_0} + \beta_2\sigma^2_{V_0} + \gamma\beta_3\sigma_{A_0}\sigma_\xi - \beta_1\sigma^2_{Y_0}}{\sigma^2_{A_0}} + \beta_3\sigma^2_\zeta\right)$
	K	Constant	K	$\frac{1}{\sqrt{\sigma^2_{\dot{z}}\sigma^2_w}}\left(\gamma\sigma_{A_0}\sigma_\zeta(\beta_1 - K)\right)$	$\frac{1}{\sqrt{\sigma^2_{\dot{z}}\sigma^2_w}}(\gamma\sigma_{A_0}\sigma_\zeta(\beta_1 - K) + \beta_3\sigma^2_\zeta)$

$\rho_{\dot{z}w}$ is obtained by substituting K in equation 8.

$$\sigma_{\bar{z}w} = E(A_t\xi) - KE(A_0\xi)$$

$$= \beta_3\sigma^2_\xi + \text{Cov } A_0\xi(\beta_1 - K)$$

$$= \beta_3\sigma^2_\xi + \gamma\sigma_{A_0}\sigma_\xi(\beta_1 - K)$$

Then

(8) $$\rho_{\bar{z}w} = \frac{\beta_3\sigma^2{}_\xi + \gamma\sigma_{A_0}\sigma_\xi(\beta_1 - K)}{\sigma_{\bar{z}}\sigma_w}$$

Equation (8) indicates that if

1. $\gamma = 0$, (i.e., no initial confounding) and/or
2. $\beta_1 = K$,

the statement $\rho_{\bar{z}w} = 0$ is equivalent to $\beta_3 = 0$ (provided that the variances are all greater than zero.)

Defining Values of K

In practice, the regression coefficient for predicting $y_{(t)}$ from $y_{(0)}$ is the value most frequently chosen to represent K. Because of the nested nature of the data, however, there are three such regression coefficients. In order to examine the appropriateness of using any one of these coefficients for K, the relationships between each of the coefficients and the structural coefficients are derived and shown in the following section.

Relationships Between Regression Coefficients and the Structural Model

The total regression coefficient (β_T), the between regression coefficient (β_B) and the within regression coefficient (β_W) can be expressed in terms of the model components as follows (Table 2);

By definition,

$$\beta_T = \frac{EE_j(Y(t)_{ij} - M_{Y(t)})\,(Y(0)_{ij} - M_Y(0))}{E(Y(0)_{ij} - M_Y(0))^2}$$

As before, both $M_{y(t)}$ and $M_{y(0)}$ are zeros.

The numerator is $\mathrm{Cov}(Y_{(t)ij}, Y_{(0)ij})$ and the denominator is $\mathrm{Var}(Y_{(0)ij})$

Thus,

$$\beta_T = \frac{\mathrm{Cov}(Y(t)_{ij}\ ,\ Y(0)_{ij})}{\mathrm{Var}(Y(0)_{ij})}$$

Substituting $\mathrm{Cov}(Y_{(t)ij}, Y_{(0)ij})$ and $\mathrm{Var}(Y_{(0)ij}$ for their corresponding values in Table 1 yields

$$\beta_T = (\beta_1\sigma^2 A_0 + \beta_2\sigma^2 V_0 + \beta_3\gamma\sigma_{A_0}\sigma_\xi)/(\sigma^2 A_0 + \sigma^2 V_0 + \sigma^2 e_0)$$

Similiarly, the between regression coefficient is,

$$\beta_B = \frac{E(\bar{Y}(t)_j - M_{Y(t)})\,(\bar{Y}(0)_j - M_{Y(0)})}{E(\bar{Y}(0)_j - M_{y(0)})^2}$$

$$= \mathrm{Cov}\ (\bar{Y}_{(t)j}, \bar{Y}(0)_j)\ /\ \mathrm{var}\ (\bar{Y}_{(0)j})$$

By using equation 4 to obtain the means of $Y_{(t)ij}$, $Y_{(0)ij}$ and by substitution

$$\beta_B = \frac{\frac{1}{s}\beta_2\sigma^2{}_{V_0} + \beta_1\sigma^2{}_{A_0} + \gamma\beta_3\sigma_{A_0}\sigma_\xi}{\sigma^2{}_{A_0} + \frac{1}{s}\sigma^2{}_{V_0} + \frac{1}{s}\sigma^2{}_{e_0}}$$

Similiarly, the within regression coefficient is,

$$\beta_w = \frac{E_j(Y(t)_{ij} - M_{Y(t)_j})(Y(0)_{ij} - M_{Y(0)_j})}{E_j(Y(0)_{ij} - M_{y(0)_j})^2}$$

$$= \beta_2(\sigma^2{}_{V_0}\ /\ (\sigma^2{}_{V_0} + \sigma^2{}_{e_0}))$$

As shown in Table 2, when $\beta_3 = 0$, $\rho_{\bar{z}w}$ equals zero if

(9) $\quad \gamma\ ((\beta_1 - \beta_2)\sigma^2{}_{V_0} + \beta_1\sigma^2{}_{e_0}) = 0$

irrespective of the choice of regression coefficients. for ρ_{zw} = to be equivalent to $\beta_3 = 0$, equation 9 is both the necessary and sufficient condition.

Conditions under which $\gamma((\beta_1 - \beta_2)\sigma^2_{V_0} + \beta_1\sigma^2_{e_0}) = 0$

<u>When $\gamma = 0$</u>

If $\gamma = 0$, irrespective of the relationships among β_1, β_2, $\sigma^2_{V_0}$ and $\sigma^2_{e_0}$ or the choice of K, equation 9 will be true. Put another way, when $\gamma = 0$, there is no problem of adjusting the achievement criterion for initial confounding with the teacher behavior.

<u>When $\gamma \neq 0$</u>

If γ does not equal zero, for equation 9 to hold, $(\beta_1 - \beta_2)\,\sigma^2_{V_0} + \beta_1\sigma^2_{e_0}$ must equal zero. This can happen when

1. $\beta_1 = \beta_2\,(\sigma^2_{V_0} / (\sigma^2_{V_0} + \sigma^2_{e_0})$ or,

2. $\beta_1 = \beta_2$ and $\sigma^2_{e_0} = 0$

The former can happen only under unlikely circumstance. The latter can happen, if a perfectly reliable premeasure is used (so that $\sigma^2_{e_0} = 0$), and when subjects are randomly assigned to clasrooms (so that $\beta_1 = \beta_2$).

In examing the relationship between $\rho_{\bar{z}w}$ and β_3, none of the conditions identified seems likely to obtain in practice. Random assignment can rarely be achieved in practice and perfectly reliable achievement measures rarely exist.

An alternative to using a regression coefficient as a method for defining K would be to estimate β_1, directly. For example, from Table 2

$$\beta_T = \frac{\beta_2\sigma^2_{V_0} + \beta_1\sigma^2_{A_0}}{\sigma^2_{A_0} + \sigma^2_{A_0} + \sigma^2_{e_0}} \quad \text{when } \beta_3 = 0 \quad \text{thus, } \beta_1 = \frac{\beta_T\sigma^2_{Y_0} - \beta_2\sigma^2_{V_0}}{\sigma^2_{A_0}}$$

Since, $E(MS_{\beta_{y_0}}) = s\sigma^2_{A_0} + \sigma^2_{V_0} + \sigma^2_{e_0}$, and

$$E(MS_{w_{y_0}}) = \sigma^2_{V_0} + \sigma^2_{e_0}$$

$$\sigma^2_{A_0} = \frac{1}{s}(MS_{\beta_{y_0}} - MS_{w_{y_0}})$$

From Table 1

$$\beta_w = \beta_2 (\sigma^2_{V_0} / (\sigma^2_{V_0} + \sigma^2_{e_0}))$$

$$\hat{\beta}_2 \sigma^2_{V_0} = \hat{\beta}_w MS_{w_{y_0}}$$

$$\hat{\beta}_1 = (\hat{\beta}_T((Ms_B - MS_W) / s + MS_W) - \hat{\beta}_W MS_W) / \ (MS_B - MS_W) / s \ )$$

or $\hat{\beta}_1 = \hat{\beta}_T \ MS_B + ((s-1)\hat{\beta}_T - S\hat{\beta}_W)MS_W \ / \ ((MS_B - MS_W)$

Distributions of Test Statistics

Even under conditions where if $\beta_3 = 0$ then $\rho_{\bar{z}w} = 0$, a t-test of $\rho_{\bar{z}w} = 0$ could still be inappropriate due to the effect of having estimated the value of K based on sample data rather than setting K a priori to a known constant. Thus what remains to be done is to determine the effects on the distribution of "t" due to estimating K in each of the several ways. The "t" distribution of $r_{\bar{z}w}$ is defined as the sampling distributions of the t ratio with c-2 degrees of freedom which is obtained from $r_{\bar{z}w}$ using the equation

(10)

$$t = \frac{r_{\bar{z}w} \sqrt{c-2}}{\sqrt{1 - r^2_{\bar{z}w}}}$$ (Hays, 1973, p. 661).

Since the exact nature of the "t" distributions of $r_{\bar{z}W}$'s could not be determined, a simulation study was conducted. In addition to using estimates of β_T, β_B and β_W to form residual gain score, Z_1, Z_2 and Z_3 respectively, the use of the proposed estimate of β_1, was used to form residual gain score Z_4. For comparison, another form of residual gain score, Z_5, was formed by setting K to a priori known constant (i.e., $K=\beta_1$).

CHAPTER IV

SIMULATION PROCEDURE

As shown in the previous chapter, testing $\rho_{\bar{z}w} = 0$ is equivalent to testing H_0: $\beta_3 = 0$ if either of the following conditions are met; 1) $\gamma = 0$, 2) $\beta_1 = \beta_2$, given a perfectly reliable premeasure. Interestingly it was found that for both of these two situations the equivalence between $\rho_{\bar{z}w} = 0$ and $\beta_3 = 0$ is true regardless of whether Z is defined using K set to the total, between or within regression coefficient or any other values of K for that matter. However, in practice, the parametric values of β_T, β_B, β_W and β_1 are seldom known. Thus, the purpose of the simulation study was to investigate the appropriateness of using a t-test to test $\rho_{\bar{z}w} = 0$ in situations where estimates are used for β_T, β_B, β_W and β_1. The empirical sampling distribution of "t" statistics for each of the four methods of defining residual gain scores were simulated and compared with the central t-distribution. The means of the empirical sampling distributions, empirical Type I error rates, and empirical powers were used to determine the appropriateness of using a t-test to test $\rho_{\bar{z}w}=0$.

The procedures employed in this empirical study will now be discussed. First, the description of the simulation parameters will be given, and then the data generation routine will be described.

Simulation Parameters

As stated previously, this investigation required the study of random sampling distributions of "t" based on $r_{\bar{z}_1 w}$, $r_{\bar{z}_2 w}$, $r_{\bar{z}_3 w}$, $r_{\bar{z}_4 w}$ and $r_{\bar{z}_5 w}$.

Empirical generation of the random sampling distributions was done repeatedly taking random samples from a known population, an approach which is typically referred to as Monte Carlo. The parameters of interest were the number of classes per sample, the number of students withing each class, the value of β_1 relative to β_2, the reliability of the premeasure, the magnitude of initial confounding, and the central and non-central cases.

As previously stated, the means of the manifest variables, Y_t, Y_0 and W were set equal to zero. Also, without loss of generality $\sigma^2 y_t$, $\sigma^2 y_0$ and $\sigma^2 w$ were set equal to 1. Y_t, Y_0 and W were assumed to have a multivariate normal distribution.

Both the number of classes, c, and the number of students per class, s, were allowed to vary so that effects on the distributions of the various "t" statistics could be investigated. The number of classes was set at 10, 30 and 50. Ten classes (or teachers) were chosen as an easily obtainable sample size. Fifty classrooms were chosen as an unusually large sample size. The number of students per class was set at 10, 20 and 30. The size of 10 was chosen as a lower bound for classroom size which might occur through loss of data. Class sizes of 20 and 30 are typical of schools today.

While β_2 represents the within class regression slope, given a perfect premeasure, β_1 does not represent exactly the between slope, as shown in Table 2. Consequently the exact magnitude of β_1, relative to β_2 cannot be decided. Therefore, three different combinations of β_1 and β_2 were selected. First, β_1 was set equal to β_2 with value equal to .7. Second, β_1 was set greater than β_2 with values .7 and .3 respectively. Third, β_1 was set smaller than β_2 with values .3 and .7 respectively. The last situation was included for comparison in spite of the fact that it is rarely encountered in practice. (e.g., Cronbach, 1976)

Table 3

Design of Study

Reliability	Teacher Behavior Effect	β_1 vs β_2	$\gamma = 0$									$\gamma = .2$									$\gamma = .4$								
		# of Classes	10			30			50			10			30			50			10			30			50		
		# of Students within Class	10	20	30	10	20	30	10	20	30	10	20	30	10	20	30	10	20	30	10	20	30	10	20	30	10	20	30
$\rho_{y_oy_o} = 1$	$\beta_3 = 0$	$\beta_1=\beta_2$														*													
		$\beta_1<\beta_2$														*													
		$\beta_1>\beta_2$														*													
	$\beta_3 = .1$	$\beta_1=\beta_2$														*													
		$\beta_1<\beta_2$														*													
		$\beta_1>\beta_2$														*													
$\rho_{y_oy_o} = .8$	$\beta_3 = 0$	$\beta_1=\beta_2$					*						*		*	*	*		*						*				
		$\beta_1<\beta_2$					*						*		*	*	*		*						*				
		$\beta_1>\beta_2$					*						*		*	*	*		*						*				
	$\beta_3 = .1$	$\beta_1=\beta_2$					*						*		*	*	*		*						*				
		$\beta_1<\beta_2$					*						*		*	*	*		*						*				
		$\beta_1>\beta_2$					*						*		*	*	*		*						*				

When $\beta_1 \neq \beta_2$ the ratio of between variation to within variation varies with the number of students per class. That is, the intraclass correlation, PI, gets smaller as the number of students per class increases. The intraclass correlation was set at .30 regardless of c and s for the present study. This value was chosen because there is evidence, for example in school mathematics, that actual school variation accounts for 30 percent of the student achievement of MAT mathematics scores (Haney, 1974).

Since Y_t and Y_0 contain errors of measurement, the estimators of the different adjustment coefficients (i.e., β_T, β_B and β_W) will be biased. The magnitude of bias is proportional to the reliability of Y_0. In other words, the bias depends on the premeasure only (Porter, 1971). The reliability of both pre and post measure was set to .8. This value was chosen as a moderate reliability for achievement tests (Ebel, 1979). Since measurements of teacher behavior have lower reliability (Brophy, 1974), .5 was selected as the reliability coefficient of W.

As a result of setting $\sigma^2 y_t$, $\sigma^2 Y_0$, $\sigma^2 w = 1$, $\rho\, Y_t Y_e$, $\rho\, Y_0 Y_0 = .8$ and the reliability of w to 5 the values taken by $\sigma^2 e_t$, $\sigma^2 e_0$, $\sigma^2 e_\xi$ and σ^2_ξ were .2, .2, .5 and .5, respectively. Also, as a result of setting PI = .3 the values taken by $\sigma^2 A_0$ and $\sigma^2 v_0$ were .24 and .56, respectively, in the presence of errors of measurement.

Three levels of initial confounding were considered: $\gamma = 0$ to indicate no confounding $\gamma = .4$ to indicate substantial confounding, and $\gamma = .2$ as an intermediate level of initial confounding.

Lastly, both the central and non-central cases were included in the study to examine the probability of Type I and Type II errors. For the purpose of this study, β_3 was set equal to 0.00 and 0.10. .1 was chosen as an arbitrary value to indicate the non-central case. Table 3 illustrates all possible combinations of the six design dimensions included in the simulation study. An "*" marks the

cells examined. These cells were selected to facilitate the investigation of the effects of initial confounding, presence of errors of measurement in the premeasure, relative magnitude of β_1 to β_2, sample and class sizes on the distribution of "t" statistics for different methods of defining r_{ZW}. One thousand samples were simulated for each of the selected cases.

Data Generation Routine

Three manifest variables were generated Y_t, Y_o and W. The three variables were generated to have a multivariate normal distribution with a mean vector of zero's and a variance covariance matrix (see Table 1). As shown in equations 4 and 5 in the analytic chapter, the manifest variables are defined

$$Y_t = A_t + V_t + e_t,$$

$$Y_0 = A_0 + V_0 + e_0,$$

$$W = \xi + e_{\xi}$$

where all the components have been defined previously. Thus, Σ can be decomposed into Σ_w, Σ_B and Σ_e, the within, between and errors of measurement variance covariance matrices respectively, as shown in Table 4. Having identified the set of parameters for each population, the Cholesky factor was computed for the between and within population variance-covariance matrix. These were used to transform generated between and within normal variates with (0,1) into between and within components with the desired vector of means and variance covariance matrix.

A FORTRAN program was written to generate the sample data and compute summary statistics for each sample. In order to generate the sample data, the between, within and errors of measurement components needed to be generated.

Table 4

Between (Σ_β), Within (Σ_w) and Errors of Measurement (Σ_e) Variance Covariance Matrix

	Σ_β			Σ_w			Σ_e		
	Y(t)	Y(0)	W	Y(t)	Y(0)	W	Y(t)	Y(0)	W
Y_t	$\sigma^2_{A_t}$			$\sigma^2_{V_t}$			$\sigma^2_{e_t}$		
Y_o	$\beta_1\sigma^2_{A_0} + \beta_3\gamma\sigma_{A_0}\sigma_\xi$	$\sigma^2_{A_0}$		$\beta_2\sigma^2_{V_0}$	$\sigma^2_{V_0}$		0	$\sigma^2_{e_0}$	
w	$\beta_1\gamma\sigma_{A_0}\sigma_\xi + \beta_3\sigma^2_\xi$	$\gamma\sigma_{A_0}\sigma_\xi$	σ^2_ξ	0	0	0	0	0	$\sigma^2_{e\xi}$

Concerning the between components, two basic steps were used to generate A_t, A_0 and ξ. First, a vector of independent normal variates, $\underline{L}$, was generated by calling the function GGNQF three times, once for each latent variable. This function which is adapted by IMSL (1982) generates one pseudo random normal deviate (0,1) every time it is called. Second, the obtained normal variates were transformed into a vector of A_t, A_0, ξ. This was done by multiplying $\underline{L}$ with the transpose of the cholesky factor of Σ_B(denote T'). This can be summarized as

$$\begin{bmatrix} A_t \\ A_0 \\ \xi \end{bmatrix} = T' \times \underline{L}$$

Steps one and two were repeated as many times as the number of classes in the sample, c. The obtained A_t, A_0, ξ had a multivariate normal distribution with a vector mean of zero and Σ_B variance covariance matrix. The within components V_t, V_0 were generated in a similar way as the between components except Σ_W was used instead of Σ_B.

GGNQF was also used to generate the normal deviates used to form errors of measurement for the manifest variables. The normal deviates were then, mulitiplied by the standard error of measurement.

Having generated the between, within and error components, each manifest variable was obtained by addition of its components parts.

A subroutine was written to compute the different forms of $r_{\bar{z}w}$'s. The obtained sample correlation coefficients were transformed into a t-ratio with c-2 degrees of freedom using equation 10. Throughout this dissertation the empirical t-sampling distribution of $r_{\bar{z}_1w}$ will be denoted as t_{z_1w}, $r_{\bar{z}_2w}$ as t_{z_2w}, $r_{\bar{z}_3w}$ as t_{z_3w}, $r_{\bar{z}_4w}$ as t_{z_4w} and $r_{\bar{z}_5w}$ as t_{z_5w}.

Another subroutine was written to obtain empirical Type I and Type II errors for the t_{zw}'s at nominal values of .005, .01, .025, .05, .1, .995, .99, .975, .95, and .90. This allowed consideration of fit for both one and two tailed tests of the null hypothesis $\rho_{\bar{z}w} = 0$.

In order to check the accuracy of the computer program written to calculate summary statistics, the simulated data for the 5 classes with 5 students each design were printed out and analyzed separately using the SPSS statistical package. The results of the two sets of calculation agreed perfectly. The simulation portion of the program was verified by executing the program to obtain Type I errors for a set of parameters in which Y_t, Y_o, W were perfectly reliable, $\gamma = 0$ and $\beta_1 = \beta_2$. Under these conditions the different t_{zw}'s all have a central t-distribution. The empirical Type I errors of the t_{zw}'s were in close agreement to their corresponding nominal alphas. For example, the empirical Type I errors of t_{z_1w}, t_{z_2w}, t_{z_3w}, t_{z_4w} and t_{z_5w} were .049, .049, .051, .048, .051 for upper tail $\alpha = .05$ and .054, .052, .056, .052, .050 for .05 lower tail $\alpha = .05$; .100, .096, .100, .099 and .101 for upper tail $\alpha = .10$ and .105, .107, .106, .110 and .105 for lower tail $\alpha = .10$ nominal alpha.

For each cell identified in Table 3 the program was run once. The seed number for every run was the random number generated after the last one used by the preceeding run.

CHAPTER V

RESULTS OF THE EMPIRICAL INVESTIGATION

In Chapter III, it was shown that Ho: $\rho_{\bar{z}w} = 0$ is equivalent to Ho: $\beta_3 = 0$ if either of the following conditions are met: 1) $\gamma = 0$ 2) $\beta_1 = \beta_2$ and a perfectly reliable premeasure. When these conditions are not met, however, $\rho_{\bar{z}w} = 0$ is equivalent to H_0: $\beta_3 = 0$ only for Z_4 and Z_5. This chapter demonstrates empirically the Type I error and power of this first test statistics of Ho: $\rho_{\bar{z}w}=0$ for situations which are common in educational research.

The variables of interest in the empirical investigation were: magnitude of initial confounding, reliability of the premeasure, relative magnitude of β_1 to β_2, number of classes per sample (sample size), and number of students within each class. Any combination of levels of the above variables identifies a sampling distribution for each of the several $r_{\bar{z}w}$'s. The specific sampling distributions investigated were selected according to a design which facilitated investigation of the effects of each of the several design variables while holding the other variables constant. The subset of sampling distributions chosen to study is represented by asterisks in the six dimensional matrix in Table 3.

The effects of initial confounding, presence of errors of measurement in the premeasure, sample and class sizes on the mean estimated of $\rho_{\bar{z}w}$'s, the empirical Type I errors and empirical power of the one and two tailed tests of $\rho_{\bar{z}w}$'s are presented in this chapter.

In general, the results of the study showed that when there was a substantial amount of initial confounding, the test statistics for t_{z_1w}, t_{z_2w} and t_{z_3w} were only valid in a few situations. These tests, particularly t_{z_1w} and

t_{z_3w}, tended to be too liberal in situations where $\beta_1 = \beta_2$ or $\beta_1 > \beta_2$ and too conservative when $\beta_1 < \beta_2$. Parallel results for t_{z_1w} and t_{z_3w} were obtained with increasing sample size. However, the test statistics for t_{z_4w} and t_{z_5w} were the only tests which remained valid across all levels of initial confounding, presence of errors of measurement, sample and class sizes. Furthermore, the results of the study indicated that increasing sample and class size and presence of errors of measurement increased the empirical power of both t_{z_4w} and t_{z_5w} in situations where $\beta_1 = \beta_2$ or $\beta_1 > \beta_2$.

Mean Estimates of $\rho_{\bar{z}w}$ when $\beta_3 = 0$

Initial Confounding Effects

By examining the equations in column 5 of Table 2, one can predict that when $\gamma = 0$ and $\beta_3 = 0$ each of the five $r_{\bar{z}w}$'s under investigation have expected value equal to zero. The numerators of these equations are

$\gamma\sigma_{A_0}\sigma_\xi\ ((\beta_1 - \beta_2\ \sigma^2_{V_0} + \beta_1\sigma^2_{e_0})$ for $\rho_{\bar{z}_1w}$, $\rho_{\bar{z}_2w}$, and $\rho_{\bar{z}_3w}$,

0 for $\rho_{\bar{z}_4w}$ and $\gamma\sigma_{A_0}\sigma_\xi\ (\beta_1 - K)$ for $\rho_{\bar{z}_5w}$.

Given these numerators, one can see that all $\rho_{\bar{z}w}$'s increase as γ increases, holding other variables constant. Inspection of the numerators also makes clear that the sign and magnitude of the $\rho_{\bar{z}w}$'s is affected by the relationship of β_1 to β_2. For example, when $\beta_1 > \beta_2$ and is large, the mean estimates of $\rho_{\bar{z}_1w}$, $\rho_{\bar{z}_2w}$ and $\rho_{\bar{z}_3w}$ are expected to depart positively from zero. Similarly, when $\beta_1 = \beta_2$ (and errors of measurement are present) the departure of these mean estimates will be in the positive direction but not as far as was the case when $\beta_1 > \beta_2$. In situations where $\beta_1 < \beta_2$ and γ is large, the departure of the mean estimates of $\rho_{\bar{z}_1w}$, $\rho_{\bar{z}_2w}$ and

$\rho_{\bar{z}_3w}$ will be negative given $(\beta_1-\beta_2)\sigma^2_{V_0} > \beta_1\sigma^2_{e_0}$. The mean estimate of $\rho_{\bar{z}_2w}$ is expected to be smaller in absolute value than the mean estimate of $\rho_{\bar{z}_1w}$ and $\rho_{\bar{z}_3w}$. This is because all three share the same numerator but $\rho_{\bar{z}_2w}$ has the largest denominator. The denominators, as shown in Table 2, are:

$$\sqrt{\sigma^2_{\bar{z}}\sigma^2_w}\ (\sigma^2_{A_0} + \sigma^2_{V_0} + \sigma^2_{e_0}) \text{ for } \rho_{\bar{z}_1w},$$

$$\sqrt{\sigma^2_{\bar{z}}\sigma^2_w}\ (S\sigma^2_{A_0} + \sigma^2_{V_0} + \sigma^2_{e_0}) \text{ for } \rho_{\bar{z}_2w},$$

$$\sqrt{\sigma^2_{\bar{z}}\sigma^2_w}\ (\sigma^2_{V_0} + \sigma^2_{e_0}) \text{ for } \rho_{\bar{z}_3w}.$$

In summary, given γ is large, it is predicted that in situations where $\beta_1 > \beta_2$, the empirical sampling distributions of $r_{\bar{z}_1w}$ and $r_{\bar{z}_3w}$ will be centered to the right of the central t-distribution and to its left when $\beta_1 < \beta_2$ (though these also depend on the magnitude of errors of measurement.) Also, it is expected that the empirical sampling distributions of $r_{\bar{z}_4w}$ and $r_{\bar{z}_5w}$ will be the closest to the central t-distribution across all combinations of β_1 and β_2.

Table 5 shows the effect of initial confounding on the mean estimates of $\rho_{\bar{z}_1w}$, $\rho_{\bar{z}_2w}$, $\rho_{\bar{z}_3w}$, $\rho_{\bar{z}_4w}$ and $\rho_{\bar{z}_5w}$ under the the three different combinations of β_1 and β_2, where sample size and class size were held constant at 30 and 20 respectively, and $\rho_{y_oy_o} = .8$. As expected, the means of the empirical sampling distributions of $r_{\bar{z}w}$'s were all near zero when $\gamma = 0$. As γ increased to .2 the mean estimates of $\rho_{\bar{z}_1w}$ and $\rho_{\bar{z}_3w}$ increased to .026 and .033 when $\beta_1 = \beta_2$ and .058, .07, respectively when $\beta_1 > \beta_2$. However, their values decreased to -.02 and -.028 when $\beta_1 < \beta_2$. Increasing to .4 caused the sampling distribution mean estimates of $\rho_{\bar{z}_1w}$ and $\rho_{\bar{z}_3w}$ to depart far from zero, particularly in situations where $\beta_1 > \beta_2$. Their values were .05, .066 when $\beta_1 = \beta_2$, -.047, -.059 when $\beta_1 < \beta_2$ and .119 and .147 respectively when $\beta_1 > \beta_2$.

While the sampling mean estimates of $\rho_{\bar{z}_2w}$ remained relatively close to zero across all levels of γ and across all combinations of β_1 and β_2, there was a

Table 5
Means of Empirical Sampling Distributions of $\rho_{\bar{z}w}$'s for Different Combinations of β_1, β_2 and for c = 30, s = 20, ρ_{YoYo} = .8 and β_3 = 0.00

	$\gamma = 0.0$					$\gamma = .2$					$\gamma = .4$				
	$r_{\bar{z}_1w}$	$r_{\bar{z}_2w}$	$r_{\bar{z}_3w}$	$r_{\bar{z}_4w}$	$r_{\bar{z}_5w}$	$r_{\bar{z}_1w}$	$r_{\bar{z}_2w}$	$r_{\bar{z}_3w}$	$r_{\bar{z}_4w}$	$r_{\bar{z}_5w}$	$r_{\bar{z}_1w}$	$r_{\bar{z}_2w}$	$r_{\bar{z}_3w}$	$r_{\bar{z}_4w}$	$r_{\bar{z}_5w}$
$\beta_1 = \beta_2$	.0012	.0003	-.0016	.00009	.0008	.026	.005	.033	.0003	-.0048	.05	.01	.066	.003	.025
$\beta_1 < \beta_2$	.0005	.006	.0006	.00002	.0000	-.02	-.003	-.028	.0005	-.011	-.047	-.0086	-.059	-.0025	.0019
$\beta_1 > \beta_2$	.0007	.00015	.0009	.0002	-.0023	.058	.012	.07	.0008	.0074	.119	.024	.147	.0035	.0067

slight increase in the mean of $r_{\bar{z}_2w}$'s in situations where $\beta_1 > \beta_2$ as increased. Mean $r_{\bar{z}w}$'s were .0015 at γ =0, .012 at γ =.2, and .024 at γ =.4. However, these mean estimates were close to zero only because the specific values of $\rho_{Y_0Y_0}$ and $(\beta_1 - \beta_2)$ were such that the two parts of the numerator in $\rho_{\bar{z}_2w}$ compensated each other.

The sampling mean estimates of $\rho_{\bar{z}_4w}$ and $\rho_{\bar{z}_5w}$ remained the closest to zero across all levels of γ and across all combinations of β_1 and β_2.

Effects of presence of errors of measurement ($\rho_{y_oy_o} \neq 1$)

σ^2_e is a common component shared by the numerators of $\rho_{\bar{z}_1w}$, $\rho_{\bar{z}_2w}$ and $\rho_{\bar{z}_3w}$. Since σ^2_e has a positive or zero value its presence should increase the departure of mean estimates of $\rho_{\bar{z}_1w}$, $\rho_{\bar{z}_2w}$ and $\rho_{\bar{z}_3w}$ from zero in situation where $\beta_1 = \beta_2$ or $\beta_1 > \beta_2$. However, this departure decreases in situations where $\beta_1 < \beta_2$. Due to the absence of σ^2_e from the equations of $\rho_{\bar{z}_4w}$ and $\rho_{\bar{z}_5w}$, errors of measurement were expected to have no effect on their sampling mean estimates.

Table 6 reports the effect of the presence of errors of measurement in the premeasure on the sampling mean estimates of $\rho_{\bar{z}w}$'s for the three different combinations of β_1 and β_2 for c = 30, s = 20 and γ=.2.

As expected, the mean estimates of $\rho_{\bar{z}_1w}$, $\rho_{\bar{z}_2w}$ and $\rho_{\bar{z}_3w}$ increased due to presence of errors of measurement when $\beta_1 = \beta_2$. While their values were all equal to .003 when $\rho_{Y_0Y_0}$ = 1.0 they became .026, .005 and .033, respectively when $\rho_{y_oy_o}$ = .8. Also, as expected, presence of errors of measurement brought the mean estimates of $\rho_{\bar{z}_1w}$ and $\rho_{\bar{z}_3w}$ closer to zero in situations where $\beta_1 < \beta_2$. Their values were -.041, -.056 when $\rho_{y_oy_o}$ = 1.0 and became -.02 and -.028 when $\rho_{y_oy_o}$ = .8. However, in situations where $\beta_1 > \beta_2$ and $\rho_{y_oy_o}$ = .8, the mean

Table 6
Means of Empirical Sampling Distributions of $\rho_{\hat{z}w}$'s for Different Combinations of β_1, β_2 and for c = 30, s = 20, γ = .2 and β_3 = 0

	$\rho_{y_0y_0}$ = 1.0					$\rho_{y_0y_0}$ = .8				
	$r_{\hat{z}_1w}$	$r_{\hat{z}_2w}$	$r_{\hat{z}_3w}$	$r_{\hat{z}_4w}$	$r_{\hat{z}_5w}$	$r_{\hat{z}_1w}$	$r_{\hat{z}_2w}$	$r_{\hat{z}_3w}$	$r_{\hat{z}_4w}$	$r_{\hat{z}_5w}$
$\beta_1 = \beta_2$	.003	.003	.003	0.001	.003	.026	.005	.033	.0003	-.0048
$\beta_1 < \beta_2$	-.041	0.00	-.056	.006	.005	-.02	-.003	-.028	-.0005	-.011
$\beta_1 > \beta_2$	.057	.009	.075	.001	-.0006	.058	.012	.07	.0008	.0074

estimates of $\rho_{\bar{z}_1w}$ and $\rho_{\bar{z}_3w}$ did not increase as expected. Their values were .057, .075 when $\rho_{y_oy_o}$ = 1 and .058, .07 when $\rho_{y_oy_o}$ = .8.

The mean estimates of $\rho_{\bar{z}_4w}$ and $\rho_{\bar{z}_5w}$ remained the closest to zero in the presence and absence of errors of measurement and across all combinations of β_1 and β_2.

<u>Sample and Class Size Effect</u>

Due to presence of s in its denominator, $\rho_{\bar{z}_2w}$ was not only expected to have a smaller mean estimate than $\rho_{\bar{z}_1w}$ and $\rho_{\bar{z}_3w}$ but also it was expected to get smaller as s increased. c is not part of any of the equations of $\rho_{\bar{z}w}$; therefore, it was expected the mean estimates of $\rho_{\bar{z}w}$'s would not be affected by changing sample size.

Table 7 shows the mean estimates of $\rho_{\bar{z}w}$'s across different levels of sample size where γ = .2, $\rho_{y_oy_o}$ = .8 and s = 20.

As expected, the mean estimates of all $\rho_{\bar{z}w}$'s were not affected by increasing c across combinations of β_1 and β_2. For example, the mean estimates of $\rho_{\bar{z}_1w}$, $\rho_{\bar{z}_2w}$, $\rho_{\bar{z}_3w}$, $\rho_{\bar{z}_4w}$ and $\rho_{\bar{z}_5w}$ were .06, -.01, .07, -.0045, .0073 for c = 10, .58, .012, .07, .0008, .0074 for c = 20 and .053, .01, .07, .0006, and .0012 for c = 50 in situations where $\beta_1 > \beta_2$. Table 8 shows the mean estimates of ρ_{zw}'s across different levels of class size where γ = .2, $\rho_{y_oy_o}$ = .8, c = 30 and $_3$ = 0.

As expected, the mean estimates of $\rho_{\bar{z}_1w}$, $\rho_{\bar{z}_3w}$, $\rho_{\bar{z}_4w}$ and $\rho_{\bar{z}_5w}$ were not affected by increasing s. The mean estimate of $\rho_{\bar{z}_2w}$ decreased slightly as s increased. For example, the mean estimates of $\rho_{\bar{z}_2w}$ were .018 for s = 10, .012 for s = 20 and .0048 for s = 30 in situations where $\beta_1 > \beta_2$.

Empirical Type I Errors for One and Two Tailed t-Tests When Testing H_o: $\rho_{\bar{z}w} = 0$

To evaluate the validity of the t-test in testing H_o: $\rho_{\bar{z}w} = 0$, the empirical values of the tests for t_{zw}'s were compared to the critical values obtained from

Table 7

Effects of Sample Size on Mean Estimates of ρ_{zw}'s for $\gamma = .2$

$\rho_{y_0 y_0} = .8$, s = 20, and $\beta_3 = 0$

	$\beta_1 = \beta_2$					$\beta_1 < \beta_2$					$\beta_1 > \beta_2$				
	$r_{\dot{z}_1 w}$	$r_{\dot{z}_2 w}$	$r_{\dot{z}_3 w}$	$r_{\dot{z}_4 w}$	$r_{\dot{z}_5 w}$	$r_{\dot{z}_1 w}$	$r_{\dot{z}_2 w}$	$r_{\dot{z}_3 w}$	$r_{\dot{z}_4 w}$	$r_{\dot{z}_5 w}$	$r_{\dot{z}_1 w}$	$r_{\dot{z}_2 w}$	$r_{\dot{z}_3 w}$	$r_{\dot{z}_4 w}$	$r_{\dot{z}_5 w}$
10	.025	.0005	.0315	-.0018	.004	-.009	.01	-.015	.0004	.01	.06	-.01	.07	-.0045	.0073
30	.026	.005	.033	.0003	-.0048	-.02	-.003	-.028	-.0005	-.011	.058	.012	.07	.0008	.0074
50	.027	.0054	.033	.005	-.001	-.023	-.004	-.024	-.0003	-.0022	.053	.01	.07	.0006	.0012

Table 8
Effects of Class Size on Mean Estimates of $\rho_{\bar{z}w}$'s for $\gamma = .2$, $\rho_{YoYo} = .8$, $c = 30$ and $\beta_3 = 0$

S	$\beta_1 = \beta_2$					$\beta_1 < \beta_2$					$\beta_1 > \beta_2$				
	$r_{\bar{z}_1w}$	$r_{\bar{z}_2w}$	$r_{\bar{z}_3w}$	$r_{\bar{z}_4w}$	$r_{\bar{z}_5w}$	$r_{\bar{z}_1w}$	$r_{\bar{z}_2w}$	$r_{\bar{z}_3w}$	$r_{\bar{z}_4w}$	$r_{\bar{z}_5w}$	$r_{\bar{z}_1w}$	$r_{\bar{z}_2w}$	$r_{\bar{z}_3w}$	$r_{\bar{z}_4w}$	$r_{\bar{z}_5w}$
10	.023	.007	.039	-.001	.0022	-.02	-.0067	.027	.0004	-.041	.005	.018	.07	-.003	.0056
20	.026	.005	.0033	.0003	-.0048	-.02	-.003	.028	-.0005	.011	.058	.012	.07	.0008	.0074
30	.026	.0019	.033	.007	-.0036	-.024	-.0029	-.029	-.0013	.005	.016	.0098	.07	.002	.0091

the t-distribution with c-2 degrees of freedom for selected level of significance. When the null hypothesis is true (i.e., $\beta_3 = 0$), the observed relative frequency of data sets having values of t_{z_1w}, t_{z_2w}, t_{z_3w}, t_{z_4w}, and t_{z_5w} greater than the critical values in the upper tail or smaller than the same critical values in the lower tail, yield the empirical levels of significance. Comparison to the selected or nominal levels of significance gives an indication of whether the test used is conservative, liberal, or correct. Comparisons were made at three nominal levels of significance which are commonly used by educational researchers; .01, .05 and .1. Observed levels of significance were in all cases based on calculating t_{zw}'s for 1000 replications from a multivariate normal distribution with specified characteristics. To facilitate comparison of empirical and nominal levels of significance, 95% probability intervals were computed using the normal approximation of the binomial distribution with n=1000 and P equal to the selected levels of significance. Thus, if the selected level of significance was .05, the 95% probability interval would be $.05 \neq 1.96\ ((.05)\ (1-.05)/(1000))^{1/2} = .05 \pm .014$. The probability limits for the nominal alpha's are presented in Tables 9 through 16. If the empirical Type I errors exceeded the upper value of the probability limit this indicated a liberal test. On the other hand, if it was less than the lower value of the probability limit this indicated a conservative test, otherwise the t-test was considered valid. The .05 nominal alpha will be chosen through out this chapter as the primary base for comparison of the different situations.

Initial Confounding Effect

It was argued earlier in this chapter, given γ is large, the empirical sampling distributions of $r_{\bar{z}_1w}$, $r_{\bar{z}_2w}$ and $r_{\bar{z}_3w}$ will be located to the right of the central t-distribution in situations where $\beta_1 = \beta_2$ or $\beta_1 > \beta_2$, and to its left

when $\beta_1 < \beta_2$. As mentioned earlier, this prediction did not hold for the sampling distribution of $r_{\bar{z}_2w}$. Also, it was argued that the empirical sampling distributions of $r_{\bar{z}_4w}$ and $r_{\bar{z}_5w}$ would be the closest to zero. As a consequence, given γ is large it was expected that using the test statistics t_{z_1w} and t_{z_3w} to test $\rho_{\bar{z}w} = 0$ would result in Liberal tests in situations where $\beta_1 = \beta_2$ or $\beta_1 > \beta_2$, and in conservative tests when $\beta_1 < \beta_2$. However, both t_{z_4w} and t_{z_5w} were expected to result in a valid test of the hypothesis of interest.

Table 9 shows the empirical Type I errors of the one tailed test of $\rho_{\bar{z}w}$ across three levels of initial confounding and across three combinations of β_1 and β_2 for c = 30, s = 20 and $\rho_{y_oy_o}$ = .8 Comparable results for the two tailed tests are shown in Table 10. It should be mentioned that here and throughout this paper, only the positive tail was considered for the one tailed tests.

$\underline{\beta_1 = \beta_2}$

All the empirical Type I errors of the one-tailed tests for t_{zw}'s were within 1.96 standard errors of their corresponding nominal alphas when $\gamma = 0$ and $\gamma = .2$. As γ increased to .4, most of the empirical Type I errors for the one-tailed tests for t_{z_1w} and t_{z_3w} were, as expected, greater than the upper limits of their corresponding nominal alphas. The other t_{zw}'s were not affected. For example, at .05 level of significance, the empirical Type I errors for one-tailed tests for t_{z_1w}, t_{z_2w}, t_{z_3w}, t_{z_4w}, and t_{z_5w} were .082, .048, .097, .043 and .047 respectively.

While the empirical Type I errors of the two-tailed tests for t_{zw}'s were in close agreement with the one-tailed t_{zw}'s when $\gamma = 0$ and $\gamma = .2$, they differed as γ increased to .4. For example, at .05 nominal alpha, the empirical Type I errors of the two-tailed tests for t_{z_1w}, t_{z_2w}, t_{z_3w}, t_{z_4w} and t_{z_5w} were .043, .029, .056, .038 and .041, respectively (Table 10). The two-tailed t_{z_1w} and t_{z_3w} were only valid due to compensating lack of fit in in each tail. Thus for $r_{\bar{z}_1w}$, $r_{\bar{z}_2w}$

Table 9
Effects of Initial Confounding on Empirical Type I Errors for the One-Tailed Tests of $\rho_{\bar{z}w}$'s = 0 where C = 30, S = 20 and $\rho_{Y_oY_o}$ = .8

β_1, β_2	Nominal Alpha and Probability Limits	$\gamma = 0$					$\gamma = .2$					$\gamma = .4$				
		$t_{/1}w$	$t_{/2}w$	$t_{/3}w$	$t_{/4}w$	$t_{/5}w$	$t_{/1}w$	$t_{/2}w$	$t_{/3}w$	$t_{/4}w$	$t_{/5}w$	$t_{/1}w$	$t_{/2}w$	$t_{/3}w$	$t_{/4}w$	$t_{/5}w$
β_1 β_2	.01 .002-.018	.012	.014	.011	.014	.014	.005	.008	.006	.008	.007	.016	.004	.02*	.006	.006
	.05 .036-.064	.052	.058	.053	.051	.058	.059	.042	.060	.039	.045	.082*	.048	.097*	.043	.047
	.1 .088-.112	.109	.105	.108	.102	.105	.109	.091	.115	.093	.095	.151*	.101	.18*	.093	.100
$\beta_1 < \beta_2$	.01 .002-.018	.009	.009	.009	.009	.009	.007	.01	.008	.01	.007	.01	.007	.01	.007	.005
	.05 .036-.064	.047	.05	.045	.045	.049	.036	.043	.034°	.044	.048	.019°	.034°	.016°	.037	.040
	.1 .088-.112	.094	.104	.098	.105	.101	.075°	.087°	.074°	.091	.106	.042°	.082°	.034°	.092	.103
$\beta_1 > \beta_2$	.01 .002-.018	.01	.01	.009	.01	.011	.02*	.012	.022*	.01	.011	.04*	.014	.058*	.011	.01
	.05 .036-.064	.055	.056	.056	.053	.058	.091*	.051	.103*	.048	.048	.148*	.083	.193*	.050	.057
	.1 .088-.112	.112	.107	.112	.105	.109	.171*	.114*	.198*	.105	.102	.26*	.127*	.304*	.105	.107

* greater than the upper limit of its corresponding nominal alpha
° smaller than the lower limit of its corresponding nominal alpha

Table 10

Effects of Initial Confounding on Empirical Type I Errors for the Two-Tailed Tests of $\rho_{\bar{z}w}$'s = 0 where C = 30, S = 20, and $\rho_{Y_oY_o}$ = .8

β_1, β_2	Nominal Alpha and Probability Limits	$\gamma = 0$					$\gamma = .2$					$\gamma = .4$				
		$t/_1w$	$t/_2w$	$t/_3w$	$t/_4w$	$t/_5w$	$t/_1w$	$t/_2w$	$t/_3w$	$t/_4w$	$t/_5w$	$t/_1w$	$t/_2w$	$t/_3w$	$t/_4w$	$t/_5w$
$\beta_1 = \beta_2$	.01 .002-.018	.007	.005	.007	.005	.005	.008	.007	.006	.009	.01	.007	.002	.011	.002	.004
	.05 .036-.064	.054	.055	.050	.057	.047	.048	.049	.05	.045	.047	.043	.029°	.056	.038	.041
	.1 .088-.112	.105	.107	.106	.112	.112	.096	.084°	.097	.084°	.095	.103	.086°	.117	.088	.103
$\beta_1 < \beta_2$	.01 .002-.018	.012	.011	.012	.011	.011	.018	.016	.018	.016	.016	.011	.003	.013	.006	.005
	.05 .036-.064	.05	.039	.047	.054	.057	.056	.05	.062	.053	.053	.043	.035°	.056	.038	.039
	.1 .088-.112	.098	.107	.098	.103	.098	.106	.10	.113*	.098	.103	.095	.080°	.097	.089	.092
$\beta_1 > \beta_2$	.01 .002-.018	.013	.012	.014	.012	.009	.012	.010	.015	.009	.007	.026*	.013	.032*	.009	.014
	.05 .036-.064	.044	.044	.044	.043	.047	.053	.091	.063	.043	.043	.094*	.040	.116*	.042	.052
	.1 .088-.112	.098	.104	.096	.101	.098	.106	.091	.117*	.091	.096	.163*	.098	.207*	.092	.105

* greater than the upper limit of its corresponding nominal alpha

° smaller than the lower limit of its corresponding nominal alpha

and $r_{\dot{z}_3w}$ either the one -tailed test was too liberal or the two-tailed test was too conservative. In contrast, both the one and two-tailed tests for t_{z_4w} and t_{z_5w} were valid in testing the hypothesis of interest across all levels of γ.

$\underline{\beta_1 < \beta_2}$

As expected, all the empirical Type I errors of the one-tailed tests were within 1.96 standard errors of their corresponding nominal alphas when $\gamma = 0$. As γ increased to .2, the empirical Type I errors for the one-tailed tests for t_{z_1w} and t_{z_3w} became slightly conservative (e.g. .036 and .034 for nominal alpha of .05) as γ increased to .4 the degree of conservativeness increased to .019 and .016 at .05 nominal alpha. As expected, the one-tailed test for t_{z_2w} also became conservative with increased initial confounding but less so than either $r_{\dot{z}_1w}$ or r_{z_2w}.

While the one-tailed tests using t_{z_1w}, t_{z_2w} and t_{z_3w} were conservative when $\gamma = .4$, only the two-tailed test using t_{z_2w} was conservative. Its empirical Type I errors was .035 at .05 nominal alpha. It should be mentioned that both the one and two-tailed tests using t_{z_4w} and t_{z_5w} were valid in testing H_0: $\rho_{\dot{z}w}=0$ across all levels of γ.

$\underline{\beta_1 > \beta_2}$

As expected, all the empirical Type I errors of the one-tailed tests were within 1.96 standard errors of their corresponding nominal alphas when $\gamma = 0$. As γ increased to .2 and to .4, the empirical Type I errors for the one-tailed tests using t_{z_1w} and t_{z_3w} increased to .091, .103 when $\gamma = .2$ and to .148, .193 respectively when $\gamma = .4$ at .05 nominal alpha. However, the one-tailed test using t_{z_2w} was not liberal at .05 as γ increased, but was at .1 nominal alpha (e.g., .114 when $\gamma = .2$ and .127 when $\gamma = .4$). None of the two-tailed tests were liberal when $\gamma = 0$ and $\gamma = .2$ at .05 nominal alpha. But as γ increased to .4 the two-tailed

tests using t_{z_1w} and t_{z_3w} became liberal (e.g., empirical Type I errors of .094 and .116 respectively at .05 nominal alpha).

Again the one and two-tailed tests using t_{z_4w} and t_{z_5w} were valid across all combinations of β_1 and β_2 and across all levels of γ.

Effects on Empirical Type I Errors of Test Statistics When the Presmeasure Contains Errors of Measurement

As mentioned earlier, presence of errors of measurement was expected to push the empirical t-sampling distributions of $r_{\bar{z}_1w}$, $r_{\bar{z}_2w}$ and $r_{\bar{z}_3w}$ to the right of a central t-distribution when $\beta_1 = \beta_2$ or $\beta_1 > \beta_2$. Also, errors of measurement were expected to bring the empirical sampling distributions of $r_{\bar{z}_1w}$, $r_{\bar{z}_2w}$ and $r_{\bar{z}_3w}$ closer to the central t-distribution in situations where $\beta_1 < \beta_2$.

Table 11 shows the empirical Type I errors of the one-tailed tests of $\rho_{\bar{z}w}$'s when $\rho_{y_oy_o} = 1.0$ and .8 across the three combinations of β_1 and β_2 for c = 30, s = 20 and γ=.2. Comparable results for the two-tailed tests are shown in Table 12.

$\beta_1 = \beta_2$

All the empirical Type I errors of the one and two-tailed tests using t_{zw}'s across both levels of $\rho_{y_oy_o}$ were within 1.96 standard errors of their corresponding nominal alphas accept for the two-tailed tests using t_{z_2w} and t_{z_4w} where the empirical Type I errors were conservative. In contrast to what was expected, at least .8 reliability of the pretest does not invalidate the $r_{\bar{z}_1w}$, $r_{\bar{z}_2w}$ and $r_{\bar{z}_3w}$ procedures when $\beta_1 = \beta_2$.

$\beta_1 < \beta_2$

As expected, the one-tailed tests using t_{z_1w}, t_{z_2w} and t_{z_3w} were less conservative when $\beta_1 < \beta_2$ and reliability of the premeasure was less than perfect. The empirical Type I errors were .013, .015, .026 when $\rho_{y_oy_o} = 1.0$ and .036,

Table 11

Effects of the Presence of Errors of Measurement in the Premeasure on the Empirical Type I Errors for One-tailed Tests of $\rho_{\bar{z}w}$'s = 0 where c = 30, s = 20, γ = .2

β_1, β_2	Nominal Alpha and Probability Limits	$\rho_{y_oy_o} = 1.0$					$\rho_{y_oy_o} = .8$				
		tz_1w	tz_2w	tz_3w	tz_4w	tz_5w	tz_1w	tz_2w	tz_3w	tz_4w	tz_5w
	.01 .002-.018	.013	.015	.013	.014	.013	.005	.008	.006	.008	.007
$\beta_1 = \beta_2$	.05 .036-.064	.043	.046	.046	.046	.046	.059	.042	.060	.039	.045
	.1 .088-.112	.100	.09	.100	.099	.097	.109	.091	.115	.093	.045
	.01 .002-.018	.004	.009	.003	.01	.012	.007	.013	.008	.01	.009
$\beta_1 < \beta_2$	.05 .036-.1	.013°	.015°	.026°	.056	.055	.036°	.043	.034°	.044	.048
	.1 .088-.112	.066°	.001	.054°	.108	.075°	.087°	.074°	.074°	.091	.106
	.01 .002-.018	.023*	.014	.029*	.013	.016	.02*	.012	.022*	.01	.011
$\beta_1 > \beta_2$	.05 .036-.0	.098*	.053	.117*	.047	.046	.091*	.051	.103*	.048	.048
	.1 .088-.112	.164*	.11	.194*	.101	.099	.171*	.114*	.198*	.105	.102

* Greater than the upper limit of its corresponding nominal alpha.
° Smaller than the lower limit of its corresponding nominal alpha.

Table 12

Effects of the Presence of Errors of Measurement in the Premeasure on the Empirical Type I Errors for Two-tailed Tests of $\rho_{\bar{z}w}$'s + 0, where c = 30, s = 20 and γ = .2

β_1, β_2	Nominal Alpha and Probability Limits	$\rho_{y_oy_o} = 1.0$					$\rho_{y_oy_o} = .8$				
		tz_1w	tz_2w	tz_3w	tz_4w	tz_5w	tz_1w	tz_2w	tz_3w	tz_4w	tz_5w
	.01 .002-.018	.01	.012	.01	.012	.011	.006	.007	.006	.009	.01
$\beta_1 = \beta_2$	.05 .036-.064	.052	.053	.051	.052	.051	.048	.049	.05	.045	.047
	.1 .088-.112	.088	.093	.089	.093	.089	.096	.084^o	.097	.084^o	.095
	.01 .002-.018	.013	.011	.015	.011	.012	.018*	.016	.018*	.016	.016
$\beta_1 < \beta_2$	.05 .036-. 64	.048	.045	.052	.05	.054	.056	.05	.062	.053	.053
	.1 .088-.112	.106	.091	.113*	.099	.102	.106	.10	.113*	.098	.103
	.01 .002-.018	.015*	.012	.017*	.011	.013	.012	.010	.015	.009	.007
$\beta_1 > \beta_2$	.05 .036-.064	.071*	.044	.077*	.043	.051	.053	.041	.063	.043	.043
	.1 .088-.112	.122*	.099	.138*	.098	.092	.106	.091	.117*	.091	.096

* Greater than the upper limit of its corresponding nominal alpha.
^{o} Smaller that the lower limit of its corresponding nominal alpha.

.043, and .034 when $\rho_{y_oy_o}$ = .8 at .05 nominal alpha. The empirical Type I errors for the one and two-tailed tests using t_{z_4w} and t_{z_5w} remained within 1.96 standard errors of their corresponding nominal alphas across both levels of $\rho_{y_oy_o}$. While presence of errors of measurement did not have any noticeable effect on the two-tailed tests for t_{z_2w}, t_{z_4w} and t_{z_5w}, less than perfect reliability of the premeasure appeared to make the t_{z_1w} and t_{z_3w} tests slightly too liberal for nominal .01 (e.g. the empirical Type I errors were both .018).

$\underline{\beta_1 > \beta_2}$

In contrast to what was expected, the one-tailed tests using t_{z_1w} and t_{z_3w} became less liberal in the presence of errors of measurement at .01 and .05 nominal alphas (e.g. empirical Type I errors of .098 and .117 for $\rho_{y_oy_o}$ = 1 but .091, .103, respectively for $\rho_{y_oy_o}$ = .8 at .05). The expected increased liberalness due to errors of measurement in the pretest did occur, however, for nominal alpha .1.

A similar decrease in liberalness was found for the two-tailed test using t_{z_1w} and t_{z_3w} (e.g. empirical Type I errors of .071, .077 for $\rho_{y_oy_o}$ = 1.0 but .053 and .063 for $\rho_{y_oy_o}$ = .8.

Once again the one and two-tailed tests using t_{z_4w} and t_{z_5w} were valid across all combinations of β_1 and β_2 and across both $\rho_{y_oy_o}$ and .8.

<u>Sample Size Effect</u>

It was expected that increased c should result in increased power. This should not affect Type I error rates for valid tests but should increase problems for tests that are too liberal (and may be even for tests that are too conservative).

Table 13 shows the empirical Type I errors of the one-tailed tests of $\rho_{\bar{z}w}$ across three levels of sample size and across three combinations of β_1 and β_2 for

Table 13

Effects of Sample Size on Empirical Type I Errors for One-Tailed Tests of $\rho_{\bar{z}w}$'s = 0 where s = 20, γ = .2 and ρ_{YoYo} = 8

β_1, β_2	Nominal Alpha and Probability Limits	C=10					C=30					C=50				
		tz_1w	tz_2w	tz_3w	tz_4w	tz_5w	tz_1w	tz_2w	tz_3w	tz_4w	tz_5w	tz_1w	tz_2w	tz_3w	tz_4w	tz_5w
$\beta_1 = \beta_2$	.01 .002 .018	.009	.011	.011	.01	.009	.005	.008	.006	.008	.007	.012	.011	.012	.01	.009
	.05 .036 .064	.052	.045	.051	.047	.048	.054	.042	.060	.039	.045	.071*	.058	.081*	.056	.056
	.1 .088 .112	.102	.094	.102	.088	.096	.109	.091	.115*	.093	.095	.13*	.104	.137*	.099	.099
$\beta_1 < \beta_2$	.01 .002 .018	.012	.009	.012	.008	.013	.007	.01	.008	.01	.007	.007	.009	.007	.009	.009
	.05 .036 .064	.039	.044	.039	.042	.046	.036	.043	.034°	.044	.048	.041	.055	.039	.059	.058
	.1 .088 .112	.09	.089	.107	.090	.088	.075°	.087°	.074°	.091	.106	.081°	.100	.071°	.103	.102
$\beta_1 > \beta_2$	.01 .002 .018	.012	.01	.015	.01	.01	.02*	.012	.022*	.01	.011	.029*	.012	.035*	.011	.007
	.05 .036 .064	.062	.049	.063	.045	.039	.091*	.051	.103*	.048	.048	.112*	.067*	.129*	.05	.062
	.1 .088 .112	.129*	.095	.135*	.096	.100	.171*	.119*	.198*	.105	.102	.190*	.131*	.212*	.117*	.116

* Greater than the upper limit of its corresponding nominal alpha

° Smaller than the lower limit of its corresponding nominal alpha

$\gamma = .2$, $s = 20$ and $\rho_{y_o y_o} = .8$. Comparable results for the two-tailed tests are shown in Table 14.

$\beta_1 = \beta_2$

As expected, the one-tailed tests for $t_{z_1 w}$ and $t_{z_3 w}$ became increasingly liberal as c increased (e.g., the empirical Type I errors were .052, .051 for c = 10, .054, .06 for c = 30 and .071, .081 for c = 50 at .05 nominal alpha). The other one-tailed tests remained valid across all levels of c and across all levels of nominal alpha.

All the empirical Type I errors for the two-tailed tests were within 1.96 standard errors of .05 nominal alpha indicating that sample size has little to no effect on the empirical Type I errors of the two-tailed tests when $\beta_1 = \beta_2$ at .05 nominal alpha.

$\beta_1 < \beta_2$

As c increased the one-tailed tests for $t_{z_1 w}$, $t_{z_2 w}$ and $t_{z_3 w}$ became increasingly conservative, particularly at .1 nominal alpha. While Type I errors were within acceptable bounds for c = 10, the empirical Type I errors were .075, .087, .074 for c = 30 and .081, .1 and .071 for c = 50. Increasing sample size did not, however, cause a problem for the validity of $t_{z_4 w}$ and $t_{z_5 w}$.

The effect of increasing sample size on the validity of tests $t_{z_1 w}$, $t_{z_2 w}$ and $t_{z_3 w}$ were just the opposite for two-tailed tests than for one-tailed tests. For example, the empirical Type I errors were .112, .113, .121 at .1 nominal alpha for c = 50 indicating all three tests were liberal.

$\beta_1 > \beta_2$

As expected, most of the empirical Type I errors of the one-tailed tests for $t_{z_1 w}$, $t_{z_2 w}$, and $t_{z_3 w}$ were beyond the upper probability limits of their

Table 14

Effects of Sample Size on Empirical Type I Errors for the Two-Tailed Tests of $\rho_{\bar{Z}W}$'s = 0 where s = 20, Υ = .2 and ρ_{YoYo} = .8

β_1, β_2	Nominal Alpha and Probability Limits	C=10 $t_{/1}^{w}$	C=10 $t_{/2}^{w}$	C=10 $t_{/3}^{w}$	C=10 $t_{/4}^{w}$	C=10 $t_{/5}^{w}$	C=30 $t_{/1}^{w}$	C=30 $t_{/2}^{w}$	C=30 $t_{/3}^{w}$	C=30 $t_{/4}^{w}$	C=30 $t_{/5}^{w}$	C=50 $t_{/1}^{w}$	C=50 $t_{/2}^{w}$	C=50 $t_{/3}^{w}$	C=50 $t_{/4}^{w}$	C=50 $t_{/5}^{w}$
	.01 (.002 .018)	.01	.009	.012	.009	.011	.008	.007	.006	.009	.01	.01	.009	.011	.009	.009
β_1 β_2	.05 (.036 .064)	.051	.049	.056	.049	.044	.048	.049	.05	.045	.047	.05	.05	.052	.049	.052
	.1 (.088 .112)	.091	.089	.096	.089	.097	.096	.084	.097	.084^{o}	.095	.110	.106	.116	.108	.109
	.01 (.002 .018)	.008	.005	.009	.004	.006	.018	.016	.018	.016	.016	.008	.01	.009	.011	.009
$\beta_1 < \beta_2$	.05 (.036 .064)	.046	.051	.048	.051	.043	.056	.05	.062	.053	.053	.045	.056	.054	.095	.057
	.1 (.088 .112)	.106	.099	.107	.099	.109	.106	.100	.113*	.098	.103	.112	.113*	.121*	.115*	.110
	.01 (.002 .018)	.006	.003	.006	.004	.008	.012	.01	.015	.009	.007	.016	.007	.023*	.006	.007
$\beta_1 > \beta_2$	.05 (.036 .064)	.053	.039	.061	.033^{o}	.048	.053	.041	.063	.043	.043	.073*	.057	.074*	.053	.058
	.1 (.088 .112)	.101	.088	.100	.088	.085	.106	.091	.117*	.091	.096	.142*	.126*	.152*	.121*	.128*

* Greater than the upper limit of its corresponding nominal alpha

o Smaller than the lower limit of its corresponding nominal alpha

corresponding nominal alphas for c = 30 and c =50. The liberalness of these tests was increased as c increased. For example, the empirical Type I errors of the one-tailed tests for t_{z_1w}, t_{z_2w} and t_{z_3w} were .62, .49, .063 for c = 10, .091, .051, .103 for c = 30 and .112, .067 and .129 for c = 50 at .05 nominal alpha.

All of the two-tailed tests were valid when c = 10 and 30 accept for t_{z_4w} which was conservative for c = 10 at .05 nominal alpha. As c increased to 50, the two-tailed tests for t_{z_1w} and t_{z_3w} became liberal at .05 nominal alpha (e.g. empirical Type I errors of .073 and .074). Surprisingly at .1 nominal alpha even the two-tailed test, for t_{z_4w} and t_{z_5w} became too liberal.

Class Size Effect

On a priori grounds it was difficult to predict the effect that varying class size might have on the validity of the several test statistics under investigation. As reported earlier only the formula for $\rho_{\bar{z}_2w}$ was a function of class size, s, and there it appeared in the denominator.

Table 15 reports the empirical Type I errors of the one-tailed tests for t_{zw}'s across three levels of class size and across three combinations of β_1 and β_2 for γ = .2, c = 30, and $\rho_{Y_0Y_0}$ = .8. Comparable results for the two-tailed tests are shown in Table 16.

$\beta_1 = \beta_2$

All empirical Type I errors for both one and two-tailed tests were within 1.96 standard errors of .05 nominal alpha. Further, the liberalness of t_{z_1w} and t_{z_3w} remained stable as s increased. These results indicate that increasing class size does not have an effect on the validity of the tests.

Table 15

Effects of Class Size on Empirical Type I Errors for the One-Tailed Tests of $\rho_{\bar{Z}W}$'s = 0, where c = 30, γ = .2, and ρ_{YoYo} = .8

β_1, β_2	Nominal Alpha and Probability Limits	S = 10					S = 20					S = 30				
		$t_{\bar{z}_1 w}$	$t_{\bar{z}_2 w}$	$t_{\bar{z}_3 w}$	$t_{\bar{z}_4 w}$	$t_{\bar{z}_5 w}$	$t_{\bar{z}_1 w}$	$t_{\bar{z}_2 w}$	$t_{\bar{z}_3 w}$	$t_{\bar{z}_4 w}$	$t_{\bar{z}_5 w}$	$t_{\bar{z}_1 w}$	$t_{\bar{z}_2 w}$	$t_{\bar{z}_3 w}$	$t_{\bar{z}_4 w}$	$t_{\bar{z}_5 w}$
	.01 .002-.018	.012	.013	.015	.013	.01	.005	.008	.006	.008	.007	.011	.006	.012	.006	.006
$\beta_1 = \beta_2$	.05 .036 .064	.062	.047	.065	.046	.05	.059	.042	.060	.039	.045	.053	.043	.060	.04	.038
	.1 .088 .112	.114*	.094	.123*	.088	.097	.109	.091	.115*	.093	.095	.12*	.097	.119*	.093	.088
	.01 .002 .018	.01	.01	.011	.01	.014	.007	.01	.008	.01	.007	.008	.01	.007	.01	.011
$\beta_1 < \beta_2$	.05 .036 .064	.048	.054	.045	.054	.054	.036	.043	.034°	.044	.048	.037	.046	.034°	.047	.054
	.1 .088 .112	.085°	.091	.083°	.098	.101	.075°	.087	.074°	.091	.106	.085°	.095	.075°	.098	.103
	.01 .002 .018	.022*	.015	.024*	.012	.016	.02*	.012	.022*	.01	.011	.024*	.009	.031*	.009	.012
$\beta_1 > \beta_2$	.05 .036 .064	.094*	.065*	.098*	.052	.062	.091*	.051	.103*	.048	.048	.102*	.060	.117*	.051	.057
	.1 .088-.112	.154*	.118*	.172*	.102	.112	.171*	.114*	.198*	.105	.102	.185*	.114*	.210*	.104	.107

* Greater than the upper limit of its corresponding nominal alpha

° Smaller than the lower limit of its corresponding nominal alpha

Table 16

Effects of Class Size on Empirical Type I Errors for the Two-Tailed Tests of $\rho_{\bar{Z}W}$'s = 0, where c = 30, γ = .2 and ρ_{YoYo} = .8

β_1, β_2	Nominal Alpha and Probability Limits	S = 10					S = 20					S = 30				
		tz_1w	tz_2w	tz_3w	tz_4w	tz_5w	tz_1w	tz_2w	tz_3w	tz_4w	tz_5w	tz_1w	tz_2w	tz_3w	tz_4w	tz_5w
	.01 .002 .018	.008	.008	.01	.09	.011	.008	.007	.006	.009	.01	.008	.007	.009	.007	.008
$\beta_1 = \beta_2$	.05 .036 .064	.05	.046	.047	.048	.049	.048	.049	.05	.045	.047	.044	.038	.046	.038	.042
	.1 .088 .112	.108	.098	.10	.104	.105	.096	.084°	.097	.084°	.095	.086°	.082°	.092	.079	.081°
	.01 .002 .018	.01	.008	.009	.008	.011	.018	.016	.018	.016	.016	.011	.008	.011	.008	.001
$\beta_1 < \beta_2$	.05 .036 .064	.056	.055	.055	.057	.058	.056	.05	.062	.053	.053	.049	.048	.040	.049	.046
	.1 .088 .112	.100	.093	.102	.094	.098	.106	.10	.112	.098	.103	.093	.089	.098	.099	.095
	.01 .002 .018	.011	.008	.014	.009	.011	.012	.010	.015	.009	.007	.02	.01	.017	.01	.013
$\beta_1 > \beta_2$	.05 .036 .064	.059	.05	.072*	.051	.054	.053	.041	.063	.043	.043	.070*	.049	.083*	.048	.054
	.1 .088 .112	.117*	.097	.114*	.09	.102	.106	.091	.117*	.091	.096	.131*	.103	.136*	.017	.108

* Greater than the upper limit of its corresponding nominal alpha

° Smaller than the lower limit of its corresponding nominal alpha

$\underline{\beta_1 < \beta_2}$

All one and two-tailed tests for t_{z_1w}, t_{z_2w}, t_{z_3w}, t_{z_4w} and t_{z_5w} were valid across all levels of s at .05, except the one-tailed test for t_{z_3w} when s=20 and 30 which was conservative (i.e., empirical Type I error of .34 in each case).

$\underline{\beta_1 > \beta_2}$

At .05 nominal alpha, the one-tailed test for t_{z_2w} was liberal for s=10 but became valid as s increased (e.g., Type I error of .065 for s = 10, .051 for s = 20 and .06 for s = 30). While this appears to be a positive effect of increasing class size it is important to note that the trend in changing Type I errors was not monotonic at .05 nominal alpha and was not present at other nominal alpha levels investigated.

Empirical Power

The value of 0.10 was chosen for β_3 in order to illustrate and contrast the power of the five test statistics for t_{zw}'s for testing the hypothesis H_0: $\beta_3 = 0$. The empirical powers of these test statistics were determined for the case where the significance levels were .01, .05 and 0.10. Since the results on the empirical powers of the 5 test procedures were similar for the three significance levels, only the results for .05 nominal alpha are reported.

In this section, only the tests which were found to have empirical Type I errors within two standard errors of the nominal alphas when $\rho_{\bar{z}w} = 0$ when $\beta_3 = 0$ are discussed. A "+" mark was used through the following tables to identify the tests. The power of a test is defined as the probability of correctly rejecting the null hypothesis. In general five factors affect the power of a test statistic: sample size, discrepancy between null and alternative hypothesis, error term, size of nominal alpha and whether a test is a one or two-

test. The conditions under which $r_{\bar{z}w}$'s increase or decrease can be determined by examining column 6 of Table 2. The population values of $\rho_{\bar{z}w}$'s when $\beta_3 \neq$ are determined by

$$\frac{1}{\sqrt{\sigma^2_{\bar{z}}\sigma^2_w}}\left(\left(\frac{\gamma\sigma_{A_0}\sigma_\xi((\beta_1-\beta_2)\sigma^2_{V_0}+\beta_1\sigma^2_e-\beta_3\gamma\sigma_{A_0}\sigma_\xi)}{\sigma^2_{A_0}+\sigma^2_{V_0}+\sigma^2_{e_0}}\right)+\beta_3\sigma^2\right) \text{ for } \rho_{\bar{z}_1 w},$$

$$\frac{1}{\sqrt{\sigma^2_{\bar{z}}\sigma^2_w}}\left(\left(\frac{\gamma\sigma_{A_0}\sigma_\xi((\beta_1-\beta_2)\sigma^2_{V_0}+\beta_1\sigma^2_e-\beta_3\gamma\sigma_{A_0}\sigma_\xi)}{\sigma^2_{A_0}+\sigma^2_{V_0}+\sigma^2_{e_0}}\right)+\beta_3\sigma^2_\xi\right) \text{ for } \rho_{\bar{z}_2 w},$$

$$\frac{1}{\sqrt{\sigma^2_{\bar{z}}\sigma^2_w}}\left(\left(\frac{\gamma\sigma_{A_0}\sigma_\xi((\beta_1-\beta_2)\sigma^2_{V_0}+\beta_1\sigma^2_{e_0})}{\sigma^2_{V_0}+\sigma^2_{e_0}}\right)+\beta_3\sigma^2_\xi\right) \text{ for } \rho_{\bar{z}_3 w},$$

$$\frac{1}{\sqrt{\sigma^2_{\bar{z}}\sigma^2_w}}\left(\left(\frac{\gamma\sigma_{A_0}\sigma_\xi((\beta_1\sigma^2_{A_0}+\beta_3\sigma^2_{V_0}+\gamma\beta_3\sigma_{A_0}\sigma_\xi-\beta_T\sigma^2_{Y_0})}{s\sigma^2_{A_0}}\right)+\beta_3\sigma^2_\xi\right), \text{ for}$$

$\rho_{\bar{z}_4 w}$,

$$\frac{1}{\sqrt{\sigma^2_{\bar{z}}\sigma^2_w}}\left(\gamma\sigma_{A_0}\sigma_\xi(\beta_1-k)+\beta_3\sigma^2_\xi\right) \text{ for } \rho_{\bar{z}_5 w}$$

Since s is not present in the formulas of $\rho_{\bar{z}_4 w}$ and $\rho_{\bar{z}_5 w}$, one can expect that class size will have no effect on the empirical power of the tests for $t_{z_4 w}$ and $t_{z_5 w}$. However, increases in class size is expected to drop the power of the test for $t_{z_2 w}$, particularly when $\beta_1 < \beta_2$. Again, class size is expected to have no effect on the empirical power of the tests for $t_{z_4 w}$ and $t_{z_5 w}$.

Effect of Initial Confounding

Table 17 reports the results of the effect of initial confounding on empirical power of one-tailed tests for t_{zw}'s for c = 30, s = 20, $\rho_{y_o y_o}$ = .8. Comparable results for the two-tailed tests are shown in Table 18.

$\beta_1 = \beta_2$

It is important to remember that only the one and two-tailed tests for $t_{z_4 w}$, $t_{z_5 w}$ and the one-tailed test for $t_{z_2 w}$ were valid at .05 nominal alpha.

The empirical power of the one and two-tailed tests for $t_{z_5 w}$ remained essentially constant across all levels of γ (e.g., the one-tailed test's empirical powers were .173 for γ = 0, .173 for γ = .2 and .168 for γ = .4). In contrast the empirical power of the one and two-tailed tests of $\rho_{\bar{z}_4 w}$ decreased as γ increased (e.g., for the one-tailed test the empirical powers decreased from .164 to .157 to .132 as γ increased from 0 to .2 to .4). Similarly there was evidence that the power of the one-tailed $t_{z_2 w}$ suffered with increasing γ (e.g., empirical power of .167 for γ = 0, .171 for γ = .2 and .148 for γ = .4).

$\beta_1 < \beta_2$

Earlier it was shown that only the one and two-tailed tests for $t_{z_4 w}$ and $t_{z_5 w}$ were valid given the null hypothesis.

Once again the empirical power for one and two-tailed tests for $t_{z_5 w}$ remained essentailly constant as γ increased (e.g. the empirical powers of the

Table 17

Effects of Initial Confounding on Empirical Power for One-Tailed Tests of $\rho_{\bar{Z}W}$'s = 0 where c = 30, s = 20, β_3 = .1 and ρ_{YoYo} = .8

β_1, β_2	Nominal Alpha	γ = 0					γ = 2					γ = .4				
		t_{z_1w}	t_{z_2w}	t_{z_3w}	t_{z_4w}	t_{z_5w}	t_{z_1w}	t_{z_2w}	t_{z_3w}	t_{z_4w}	t_{z_5w}	t_{z_1w}	t_{z_2w}	t_{z_3w}	t_{z_4w}	t_{z_5w}
$\beta_1 = \beta_2$	.01	.045 +	.047 +	.046 +	.047 +	.041 +	.071	.055	.073	.047 +	.056 +	.081	.036	.097	.034 +	.053 +
	.05	.107 +	.107 +	.164 +	.164 +	.173 +	.202	.171	.208	.157 +	.173 +	.242	.148	.268	.132 +	.168 +
	.1	.272	.267	.282	.258 +	.265 +	.314	.266	.331	.262 +	.275 +	.365	.244	.414	.223 +	.274 +
$\beta_1 < \beta_2$	.01	.039 +	.030 +	.041 +	.029 +	.036 +	.026	.031	.024	.031 +	.036 +	.011	.02	.009	.022 +	.028 +
	.05	.133	.135	.137	.139 +	.131 +	.100	.113	.097	.120 +	.124 +	.065	.095	.06	.10 +	.124 +
	.1	.244 +	.232 +	.242 +	.228 +	.233 +	.174	.197	.168	.20 +	.211 +	.156	.182	.146	.185 +	.213 +
$\beta_1 > \beta_2$	.01	.051 +	.047 +	.046 +	.039 +	.046 +	.069	.05 +	.073	.039 +	.049 +	.134	.053 +	.158	.046 +	.059 +
	.05	.153	.151 +	.160	.147 +	.157 +	.236	.173 +	.250	.159 +	.171 +	.327	.164 +	.377	.132 +	.173 +
	.1	.252 +	.250 +	.253 +	246 +	.251 +	.361	.283 +	.387	.257 +	.272 +	.471	.274 +	.514	.233 +	.264 +

+ Valid test (empirical Type I errors within two standard errors of the nominal alpha when β_3 = 0)

Table 18

Effects of Initial Confounding on Empirical Power for Two-Tailed Tests of $\rho_{\bar{z}w}$'s = 0 where c = 30, s = 20, β_3 = .1 and ρ_{YoYo} = .8

β_1, β_2	Nominal Alpha	$\gamma = 0$ tz_1w	tz_2w	tz_3w	tz_4w	tz_5w	$\gamma = .2$ tz_1w	tz_2w	tz_3w	tz_4w	tz_5w	$\gamma = .4$ tz_1w	tz_2w	tz_3w	tz_4w	tz_5w
	.01	.031 +	.032 +	.028 +	.032 +	.032 +	.045	.032 +	.046	.030 +	.035 +	.049	.017 +	.06	.016 +	.013 +
$\beta_1 = \beta_2$	.05	.104 +	.098 +	.101 +	.098 +	.100 +	.130	.097 +	.139	.094 +	.097 +	.159	.087 +	.189	.074 +	.105 +
	.1	.176 +	.714 +	.175 +	.171 +	.182 +	.206	.182 +	.212	.167 +	.183 +	.243	.151 +	.268	.135 +	.174 +
	.01	.018 +	.016 +	.019 +	.015 +	.018 +	.018	.020	.016	.021 +	.020 +	.004	.009	.004	.01 +	.01 +
$\beta_1 < \beta_2$	.05	.085 +	.082 +	.088 +	.078 +	.080 +	.058	.066	.058	.071 +	.082 +	.042	.056	.043	.059 +	.075 +
	.1	.148 +	.148 +	.150 +	.163 +	.144 +	.118	.125	.114	.132 +	.139 +	.09	.113	.088	.132 +	.148 +
	.01	.026 +	.026 +	.025 +	.024 +	.026 +	.044	.021	.049	.019 +	.029 +	.09	.031	.108	.025 +	.035 +
$\beta_1 > \beta_2$	.05	.100 +	.09 +	.091 +	.091 +	.089 +	.155	.102	.155	.090 +	.104 +	.252	.102	.267	.084 +	.108 +
	.1	.163 +	.156 +	.173 +	.152 +	.161 +	.241	.183	.256	.170 +	.172 +	.331	.175	.379	.146 +	.181 +

+ Valid test (empirical Type I errors within two standard errors of the nominal alpha when β_3 = 0)

one-tailed test were .131 for γ = 0 and .124 for both γ = .2 and γ = .4). And again the empirical power for one-tailed test for t_{z_4w} decreased from .139 to .120 to .10 as γ increased from 0 to .2 to .4. It should be noted that the two-tailed t_{z_4w} and t_{z_5w} were less powerful than the one-tailed tests.

$\beta_1 > \beta_2$

Only the one and two-tailed tests for t_{z_4w}, t_{z_5w} and the two-tailed test for t_{z_2w} were valid under the null hypothesis.

In contrast to previous results, the empirical power of the one and two-tailed tests for t_{z_5w} increased slightly as γ increased (e.g. empirical powers for one-tailed test of .157, .171, .173 as γ increased from 0 to .2 to .4). The empirical power of the one-tailed tests for t_{z_4w} did not have a clear relationship to γ but the two-tailed test had essentially constant power as γ increased (e.g. the empirical powers were near .09).

Similarly, the empirical power of the two-tailed tests for t_{z_2w} were not much affected by varying (e.g., empirical powers of .09, .102 and .102 as γ increased from 0 to .2 to .4).

Effects of Errors of Measurement in the Premeasure

Due to the absence of error of measurement components from the formulas of $\rho_{\bar{z}_4w}$ and $\rho_{\bar{z}_5w}$ one can expect that errors of measurement in the premeasure will not affect the empirical power of the test statistics using $r_{\bar{z}_4w}$ and $r_{\bar{z}_5w}$. Errors of measurement were expected to increase the power of the test statistics for t_{z_2w}.

Table 19 reports the results of the effect of errors of measurement on the empirical power of the several one-tailed tests for t_{zw}'s for c = 30, s = 20 and γ = .2. Comparable results for the two-tailed tests are shown in Table 20.

Table 19

Effects of Errors of Measurement in the Premeasure on Empirical Power for the One-Tailed Tests of ρ_{zw}'s = 0, where c = 30, s = 20, γ = .2, and β_3 = .1

β_1, β_2	Nominal Alpha	ρ_{yoyo} = 1.0					ρ_{yoyo} = .8				
		tz_1w	tz_2w	tz_3w	tz_4w	tz_5w	tz_1w	tz_2w	tz_3w	tz_4w	tz_5w
$\beta_1 = \beta_2$	.01	.027	.022 +	.030	.022 +	.027 +	.071	.055 +	.073	.047 +	.056 +
	.05	.11	.095 +	.114	.094 +	.117 +	.202	.171 +	.208	.157 +	.173 +
	.1	.202	.186 +	.204	.176 +	.205 +	.314	.266 +	.331	.262 +	.275 +
$\beta_1 < \beta_2$	.01	.02	.03	.012	.034 +	.034 +	.026	.031	.024	.031 +	.036 +
	.05	.089	.118	.078	.126 +	.139 +	.100	.113	.097	.120 +	.124 +
	.1	.158	.211	.136	.222 +	.236 +	.174	.197	.168	.200 +	.211 +
$\beta_1 > \beta_2$	.01	.057	.022	.071	.021 +	.024 +	.069	.050	.073	.039 +	.049 +
	.05	.212	.105	.246	.089 +	.102 +	.236	.173	.250	.159 +	.171 +
	.1	.338	.196	.382	.174 +	.187 +	.361	.283	.387	.257 +	.272 +

+ Valid test(empirical Type I errors within two standard errors of the nominal alpha when β_3 = 0)

Table 20
Effects of Errors of Measurement in the Premeasure on Empirical Power for Two-Tailed Tests of $\rho_{\bar{z}w}$'s = 0 where $c = 30$, $s = 20$, $\gamma = .2$ and $\beta_3 = .1$

β_1, β_2	Nominal Alpha	$\rho_{yoyo} = 1.0$					$\rho_{yoyo} = 8$				
		tz_1w	tz_2w	tz_3w	tz_4w	tz_5w	tz_1w	tz_2w	tz_3w	tz_4w	tz_5w
$\beta_1 = \beta_2$	.01	.014	.011 +	.019	.011 +	.018 +	.045	.033 +	.046	.030 +	.035 +
	.05	.071	.061 +	.073	.060 +	.075 +	.130	.087 +	.139	.094 +	.097 +
	.1	.13	.118 +	.136	.117 +	.142 +	.206	.182 +	.212	.167 +	.183 +
$\beta_1 < \beta_2$	.01	.009	.022	.011	.02 +	.024 +	.018	.020	.016	.021 +	.020 +
	.05	.07	.077	.059	.078 +	.089 +	.058	.066	.058	.071 +	.082 +
	.1	.117	.137	.113	.144 +	.159 +	.118	.125	.114	.132 +	.139 +
$\beta_1 > \beta_2$	.01	.03	.014	.04	.012 +	.014 +	.044	.021	.049	.019 +	.029 +
	.05	.111	.060	.153	.050 +	.058 +	.155	.102	.155	.090 +	.104 +
	.1	.216	.120	.249	.108 +	.120 +	.241	.183	.256	.170 +	.172 +

+ Valid test(empirical Type I errors within two standard errors of the nominal alpha when $\beta_3 = 0$)

$\beta_1 = \beta_2$

As seen earlier, only the one and two-tailed tests for t_{z_2w}, t_{z_4w} and t_{z_5w} were valid, given the null hypothesis. The empirical powers of the one and two-tailed tests for t_{z_2w}, t_{z_4w} and t_{z_5w} were increased in the presence of errors of measurement (e.g. empirical powers of .095, .171 for t_{z_2w}, .094, .157 for t_{z_4w}, .117, .173 for t_{z_5w}).

$\beta_1 < \beta_2$

Only the one and two-tailed test for t_{z_4w} and t_{z_5w} were valid under the null hypothesis.

The empirical power of the one and two-tailed test for t_{z_5w} decreased in the presence of measurement error (e.g., empirical power for the one-tailed was .139 for $\rho_{y_oy_o} = 1$ and .124 for $\rho_{y_oy_o} = .8$). Similarly, empirical power of t_{z_4w} decreased slightly in the presence of errors of measurement (e.g., .126 for $\rho_{y_oy_o} = 1$ and .120 for $\rho_{y_oy_o} = .8$).

$\beta_1 > \beta_2$

Earlier it was shown that only the one and two-tailed tests for t_{z_4w} and t_{z_5w} were valid given the null hypothesis.

Both empirical power of the one and two-tailed tests for t_{z_4w} and t_{z_5w} increased in the presence of errors of measurement (e.g., .089, .102 for $\rho_{y_oy_o} = 1.0$ and .159 and .171 for $\rho_{y_oy_o} = .8$).

Sample Size Effect

It was predicted that the empirical power of all the one and two-tailed tests for t_{zw}'s would increase as c increased. Table 21 gives the results of the effect of sample size on the empirical power of the one-tailed tests for t_{zw}'s for $s = 20, \gamma = .2$ and $\rho_{y_oy_o} = .8$. Comparable results for the two-tailed t_{zw}'s are shown in Table 22.

Table 21

Effects of Sample Size on Empirical Power for the One-Tailed Tests of $\rho_{\bar{z}w}$'s = 0 where s = 20, γ = .2, β_3 = .1 and ρ_{YoYo} = .8

β_1, β_2	Nominal Alpha	C = 10					C = 30					C = 50				
		tz_1^w	tz_2^w	tz_3^w	tz_4^w	tz_5^w	tz_1^w	tz_2^w	tz_3^w	tz_4^w	tz_5^w	tz_1^w	tz_2^w	tz_3^w	tz_4^w	tz_5^w
$\beta_1 = \beta_2$	.01	.029	.028 +	.032	.027 +	.030 +	.071	.055 +	.073	.047 +	.056 +	.122	.083 +	.131	.075 +	.097 +
	.05	.127	.108 +	.129	.101 +	.118 +	.202	.171 +	.208	.157 +	.173 +	.313	.253 +	.338	.245 +	.258 +
	.1	.21	.179 +	.218	.178 +	.195 +	.314	.266 +	.331	.262 +	.275 +	.451	.381 +	.475	.364 +	.378 +
$\beta_1 < \beta_2$	.01	.015	.017 +	.013	.018 +	.020 +	026	.031 +	.024	.031 +	.030 +	.032	.040 +	.031	.045 +	.049 +
	.05	.074	.082 +	.069	.084 +	.082 +	.100	.113 +	.097	.120	.124	.147	.178 +	.136	.182 +	.184 +
	.1	.143	.154 +	.142	.153 +	.155 +	.174	.197 +	.168	.20 +	.211 +	.244	.274 +	.216	.28	.291
$\beta_1 > \beta_2$	.01	.023	.014	.013	.011 +	.069 +	.069	.05	.073	.039 +	.049 +	.121	.076	.136	.058 +	.070 +
	.05	.099	.088	.108	.083 +	.082 +	.236	.173	.250	.159 +	.171 +	.353	.247	.363	.217 +	.231 +
	.1	.193	.142	.197	.145 +	.139 +	.361	.283	.387	.251 +	.272 +	.47	.355	.497	.322 +	.333 +

+ Valid test (empirical Type I errors within two standard errors of the nominal alpha when β_3 = 0

Table 22

Effects of Sample Size on Empirical Power for the Two-Tailed Tests of $\rho_{\bar{z}w}$'s = 0 where s = 20, γ = .2, β_3 = .1 and ρ_{YoYo} = .8

β_1, β_2	Nominal Alpha	C = 10					C = 30					C = 50				
		tz_1w	tz_2w	tz_3w	tz_4w	tz_5w	tz_1w	tz_2w	tz_3w	tz_4w	tz_5w	tz_1w	tz_2w	tz_3w	tz_4w	tz_5w
$\beta_1 = \beta_2$	.01	.021	.018 +	.021	.018 +	.015 +	.045	.032 +	.046	.030 +	.035 +	.079	.048 +	.086	.045 +	.057 +
	.05	.078	.064 +	.075	.061 +	.070 +	.130	.097 +	.139	.094 +	.097 +	.202	.159 +	.216	.151 +	.171 +
	.1	.161	.121 +	.144	.114 +	.134 +	.206	.182	.213	.167 +	.183 +	.315	.256	.340	.249 +	.261 +
$\beta_1 < \beta_2$	.01	.014	.013 +	.014	.012 +	.015 +	.018	.020 +	.016	.021 +	.020 +	.017	.024 +	.018	.025 +	.038 +
	.05	.056	.058 +	.058	.056 +	.062 +	.058	.066 +	.058	.071 +	.082 +	.087	.100 +	.08	.103 +	.121 +
	.1	.106	.109 +	.101	.102 +	.114 +	.118	.125 +	.114	.132 +	.139 +	.162	.142 +	.152	.195 +	.206 +
$\beta_1 > \beta_2$	.01	.015	.008	.013	.011 +	.013 +	.044	.021	.049	.019 +	.029 +	.074	.042	.081	.034 +	.036 +
	.05	.061	.057	.063	.052 +	.051 +	.155	.102	.155	.090 +	.104 +	.219	.145	.269	.119 +	.113 +
	.1	.116	.112	.121	.110 +	.106 +	.241	.183	.256	.170 +	.172 +	.354	.248	.364	.219 +	.233 +

+ Valid test (empirical Type I errors within two standard errors of the nominal alpha when β_3 = 0)

Earlier it was shown that only the one and two-tailed tests for t_{z_2w}, t_{z_4w} and t_{z_5w} were valid under the null hypothesis when $\beta_1 = \beta_2$ or $\beta_1 < \beta_2$.

The empirical power of these tests increased as c increased (e.g., the empirical power of the one-tailed tests for t_{z_2w}, t_{z_4w} and t_{z_5w} at c = 50 were 234, 243, and 214 percent of their power when c = 10 in situations where $\beta_1 = \beta_2$).

Only the one and two-tailed tests for t_{z_4w} and t_{z_5w} were valid given the null hypothesis, in situations, where $\beta_1 < \beta_2$. Again, the empirical power of these tests increased as c increased (e.g. the empirical power of the one-tailed tests for t_{z_4w} and t_{z_5w} were .83, .82 for c = 10, .159, .171 for c = 30 and .217, .231 for c = 50).

<u>Class Size Effect</u>

It was expected tnat increased class size would have no effect on the empirical power of the tests for t_{z_4w} and t_{z_5w}, but that the empirical power of the test for t_{z_2w} would drop as s increased.

Table 23 reports the results of the effect of number of students per class on the empirical power of the one-tailed tests for t_{zw}'s for c = 30, γ = .2 and $\rho_{y_oy_o}$ = .8. Comparable results for the two-tailed tests for t_{zw}'s are given in Table 24.

<u>$\beta_1 = \beta_2$</u>

As was shown earlier only the one and two-tailed tests for t_{z_2w}, t_{z_4w} and t_{z_5w} were valid, given the null hypothesis.

The changes in the empirical power of the one-tailed tests for t_{z_4w} and t_{z_5w} were not large but in each case power increased monotonicaly with s (e.g., empirical powers at s = 30 were .112 and .119 percent of the empirical powers when s = 10 and .102 and .103 percent of the empirical powers when s = 20). There was, however, no clear relationship between power and class size for t_{z_2w}

Table 23
Effects of Class Size on Empirical Power for the One-Tailed Tests of $\rho_{\bar{z}w}$'s = 0 where c = 30, γ = .2, β_3 = .1 and ρ_{YoYo} = .8

β_1, β_2	Nominal Alpha	S = 10					S = 20					S = 30				
		tz_1w	tz_2w	tz_3w	tz_4w	tz_5w	tz_1w	tz_2w	tz_3w	tz_4w	tz_5w	tz_1w	tz_2w	tz_3w	tz_4w	tz_5w
	.01	.056	.044 +	.062	.039 +	.047 +	.071	.055 +	.073	.047 +	.056 +	.061	.043 +	.066	.042 +	.045 +
$\beta_1 = \beta_2$	.05	.18	.16 +	.186	.143 +	.151 +	.202	.171 +	.208	.157 +	.173 +	.211	.164 +	.225	.160 +	.179 +
	.1	.286	.242 +	.301	.237 +	.245 +	.314	.266 +	.331	.262 +	.275 +	.336	.292 +	.348	.283 +	.294 +
	.01	.024	.024	.022	.026 +	.027 +	.026	.031	.024	.031 +	.036 +	.030	.031	.028	.032 +	.035 +
$\beta_1 < \beta_2$	.05	.101	.114	.096	.121 +	.126 +	.100	.113	.097	.120 +	.124 +	.084	.106	.084	.106 +	.112 +
	.1	.168	.142	.160	.195 +	.21 +	.174	.197	.168	.20 +	.211 +	.172	202	.163	.204 +	.215 +
	.01	.051	.035	.058	.027 +	.030 +	.069	.050	.073	.039 +	.049 +	.083	.050	.085	.047 +	.051 +
$\beta_1 > \beta_2$	.05	.205	.114	.216	.109 +	.125 +	.236	.173	.250	.159 +	.171 +	.246	.162	.265	.155 +	.164 +
	.1	.331	.249	.347	.211 +	.234 +	.361	.283	.387	.257 +	.272 +	.363	.281	.379	.271 +	.286 +

+ Valid test (empirical Type I errors within two standard errors of the nominal alpha when β_3 = 0)

Table 24

Effects of Class Size on Empirical Power for the Two-Tailed Tests of $\rho_{\bar{z}w}$'s = 0 where c = 30, γ = .2, β_3 = .1 and ρ_{YoYo} = .8

β_1, β_2	Nominal Alpha	S = 10					S = 20					S = 30				
		tz_1w	tz_2w	tz_3w	tz_4w	tz_5w	tz_1w	tz_2w	tz_3w	tz_4w	tz_5w	tz_1w	tz_2w	tz_3w	tz_4w	tz_5w
$\beta_1 = \beta_2$	.01	.034	.027 +	.041	.026 +	.026 +	.045	.032 +	.046	.030 +	.035 +	.032	.021 +	.038	.020 +	.020 +
	.05	.119	.106 +	.123	.091 +	.106 +	.130	.097 +	.139	.094 +	.097 +	.127	.103 +	.133	.100 +	.099 +
	.1	.19	.173 +	.192	.160 +	.162 +	.206	.182 +	.213	.167 +	.183 +	.219	.172 +	.235	.168 +	.188 +
$\beta_1 < \beta_2$	.01	.013	.011	.012	.012 +	.016 +	.018	.02	.016	.021 +	.020 +	.014	.018	.012	.017 +	.019 +
	.05	.071	.077	.072	.082 +	.075 +	.058	.066	.058	.071 +	.082 +	.068	.074	.069	.076 +	.080 +
	.1	.124	.137	.119	.141 +	.145 +	.118	.125	.114	.132 +	.139 +	.121	.127	.116	.126 +	.130 +
$\beta_1 > \beta_2$	.01	.032	.018	.036	.017 +	.013 +	.044	.021	.049	.019 +	.029 +	.048	.034	.054	.029 +	.030 +
	.05	.122	.083	.136	.075 +	.072 +	.055	.102	.155	.090 +	.104 +	.143	.103	.162	.095 +	.096 +
	.1	.212	.154	.225	.122 +	.134 +	.241	.183	.256	.170 +	.172 +	.252	.173	.272	.167 +	.173 +

+ Valid test (empirical Type I errors within two standard errors of the nominal alpha when β_3 = 0)

(empirical powers of .160, .171 and .164 as s increased). Similar but less pronounced relationship between power and class size were found for the two-tailed tests.

$\beta_1 < \beta_2$

Only the one and two-tailed tests for t_{z_2w}, t_{z_4w} and t_{z_5w} had empirical Type I errors within two standard errors of the nominal values when $\beta_3 = 0$.

The empirical powers of the one-tailed tests for t_{z_2w}, t_{z_4w} and t_{z_5w} tended to drop slightly as s increased, particularly from 20 to 30 (e.g., empirical powers of .114, .113, .,106 for t_{z_2w}, .121, .120, .106 for t_{z_4w} and .126, .124, .112 for t_{z_5w} as s increased from 10 to 20 to 30). However, the relationship between power and class size for the two-tailed tests for t_{z_2w}, t_{z_4w} and t_{z_5w} were not clear (e.g., empirical powers of .077, .066, .074 for t_{z_2w}, .082, .071, .076 for t_{z_4w}, .075, .82, .80 for t_{z_5w} as s increased from 10 to 20 to 30).

$\beta_1 > \beta_2$

Only the one and two-tailed tests for t_{z_4w}, t_{z_5w} had expirical Type I errors within two standard errors of the nominal alphas when $\beta_3 = 0$.

The empirical power of t_{z_4w} and t_{z_5w} tended to increase with class size though this relationship was most in evidence for one-tailed test at alpha .1 (e.g., empirical powers of .211, .257, .271 for t_{z_4w} and .234, .272 and .286 for t_{z_5w}).

CHAPTER VI

SUMMARY AND CONCLUSIONS

The purposes of this investigation were to determine the conditions under which testing for H_0: $\rho_{\bar{z}w}=0$ is equivalent to testing for no teacher behavior effect. Five different methods for defining Z were investigated under a variety of conditions defined by varying (a) the amount of initial confounding, (b) presence of errors of measurement in the premeasure, (c) sample size, (d) class size and (e) the relationship between β_1 (i.e., the structural slope of class effect at time t on class effect at time 0) and β_2 (i.e., the structural slope of within class deviation at time t on within class deviation at time 0).

A linear structural model which incorporates the hierarchical nature of the data and the possibility of measurement errors was provided in chapter three to determine analytically the conditions for which testing $\rho_{\bar{z}w} = 0$ is equivalent to testing $\beta_3 = 0$. The results showed that equivalence of the two null hypotheses does occur if either of the following conditions are met (1) $\gamma = 0$ (i.e., no initial confounding of teacher behavior and class compositions) (2) $\beta_1 = \beta_2$, given a perfectly reliable measure. Such equivalence between $\rho_{zw} = 0$ and $\beta_3 = 0$ is true regardless of whether Z is defined using K set to the total, between or within regression coefficients.

A Monte Carlo approach was taken to investigate the appropriateness of the different test statistics for t_{zw}'s in testing the hypothesis of no teacher behavior effect on student achievements. As expected, when $\gamma = 0$ and $\beta_3 = 0$ all of the mean estimates of $\rho_{\bar{z}w}$'s were near zero. Further, the empirical distributions of the "t" statistics for the different forms of $r_{\bar{z}w}$'s were close to

their theoretical t-distribution across all combinations investigated. Finally, all of the test statistics for t_{zw}'s were valid and t_{z_1w}, t_{z_3w} and t_{z_5w} had empirical power greater than t_{z_2w} and t_{z_4w}.

Increasing the amount of initial confounding, γ, caused the mean estimates of $\rho_{\dot{z}_1w}$, $\rho_{\dot{z}_2w}$ and $\rho_{\dot{z}_3w}$ to depart from zero, but did not effect the mean estimates of $\rho_{\dot{z}_4w}$ and ρ_{z_5w}. Results of the empirical Type I error rates paralleled, for the most part, the empirical results for values of $\rho_{\dot{z}w}$'s. Increasing γ caused t_{z_1w} and t_{z_3w} to be centered to the right of the theoretical t-distribution when $\beta_1 = \beta_2$ or $\beta_1 > \beta_2$ and to its left when $\beta_1 < \beta_2$. This caused the tests to be too liberal in the first two cases and too conservative in the third case for one tailed tests. For two-tailed tests, t_{z_1w} and t_{z_3w} were again too liberal when $\beta_1 > \beta_2$ but valid for the other two relationship between β_1 and β_2. Results of the empirical Type I error rates indicated that increasing γ caused the one-tailed test for t_{z_2w} to be too conservative when $\beta_1 = \beta_2$, the one and two-tailed tests to be too conservative when $\beta_1 < \beta_2$ and the one-tailed test to be too liberal when $\beta_1 > \beta_2$. The only tests for which empirical Type I error rates were not affected by increasing the amount of initial confounding were t_{z_4w} and t_{z_5w}. It can be concluded that as γ increased, only t_{z_2w}, t_{z_4w} and t_{z_5w} had empirical Type I errors within two standard errors of the nominal alphas when $\beta_1 = \beta_2$ and t_{z_4w}, t_{z_5w} for the other relationships between β_1 and β_2. However, in situations where t_{z_2w} was a valid test, it had greater power than t_{z_4w} but less than t_{z_5w}.

Errors of measurement in the premeasure caused the mean estimates of $\rho_{\dot{z}_1w}$, $\rho_{\dot{z}_2w}$ and $\rho_{\dot{z}_3w}$ to depart slightly from zero when $\beta_1 = \beta_2$ and to become closer to zero, at least for t_{z_1w} and t_{z_3w} when $\beta_1 < \beta_2$. However, errors of measurement did not effect the mean estimates of $\rho_{\dot{z}_4w}$ or $\rho_{\dot{z}_5w}$. The one-tailed

tests using t_{z_1w} and t_{z_3w} became less conservative as a result of the presence of errors of measurement when $\beta_1 < \beta_2$.

The effects of errors of measurement on the two tailed tests were not the same as those for the one-tailed tests. For example, errors of measurement brought the empirical Type I errors for t_{z_1w} and t_{z_3w} closer to the nominal alphas's when $\beta_1 > \beta_2$. Concerning the power of the tests with valid Type I errors, power for t_{z_4w} and t_{z_5w} tended to increase in the presence of errors of measurement in the premeasure when $\beta_1 = \beta_2$ or $\beta_1 > \beta_2$ but decreased when $\beta_1 < \beta_2$. Further t_{z_5w} had greater power than t_{z_4w} across all combinations of β_1 and β_2.

Sample size was found to have no effect on the mean estimates of ρ_{zw}'s across all combinations of β_1 and β_2. Increasing sample size affected empirical Type I error rates for the one-tailed tests using t_{z_1w} and t_{z_3w} (i.e., the tests were too liberal when $\beta_1 = \beta_2$ or $\beta_1 > \beta_2$ but too conservative when $\beta_1 < \beta_2$). While the results of the empirical Type I error rates for the two-tailed tests were parallel to the one-tailed tests when $\beta_1 > \beta_2$, they differed in situations where $\beta_1 = \beta_2$ or $\beta_1 > \beta_2$. Except for the one-tailed tests when $\beta_1 > \beta_2$, increasing sample size did not affect the empirical Type I error for t_{z_2w} across all combinations of β_1 and β_2 at .05 nominal alpha. Statistics t_{z_4w} and t_{z_5w} were the only tests to remain valid as sample size increased. The power of these two tests increased with sample size, and for most cases t_{z_5w} had greater power than t_{z_4w}.

Number of students within each class had no effect on the mean estimates of all ρ_{zw}'s. Also, it had no effect on the one and two-tailed test across all combinations of β_1 and β_2. In general the tests which were valid, conservative or liberal when classes were small remained so when class size increased. t_{z_4w}, t_{z_5w} and in some cases t_{z_2w} were the only tests which had empirical

Type errors within two standard errors of the nominal alpha when $\beta_3 = 0$. For these tests, the empirical power tended to increase with class size when $\beta_1 = \beta_2$ or $\beta_1 > \beta_2$ and to drop slightly when $\beta_1 < \beta_2$.

In conclusion, when students are randomly assigned to classrooms or when $\beta_1 = \beta_2$ and the premeasure has perfect reliability, testing H_o: $\rho_{\bar{z}w} = 0$ is equivalent to testing no teacher behavior effect. This equivalence is true regardless of whether Z is defined using K set to total, between or within regression weights. However, when students are not randomly assigned to classrooms (i.e.,$\gamma \neq 0$) which is typically the case in practice, the test statistics using t_{z_1w}, t_{z_2w} and t_{z_3w} were valid only in a few situations. In general these tests, particularly, t_{z_1w} and t_{z_3w} tended to be too liberal in situations where $\beta_1 > \beta_2$ (the typical case in education) and too conservative when $\beta_1 < \beta_2$. Interestingly, the only tests were valid for all conditions investigated were the tests for t_{z_4w} and t_{z_5w}. Since K is usually unknown in practice, the procedure of choice should be t_{z_4w}. In addition to being valid, it affords an estimate of K rather than requiring K to be known apriori.

Increasing sample size and presence of errors of measurement increased the empirical power of t_{z_4w} and t_{z_5w}. Their empirical power increased slightly with class size when $\beta_1 = \beta_2$ or $\beta_1 > \beta_2$ but decreased slightly when $\beta_1 < \beta_2$. While the empirical power of t_{z_5w} remained constant as γ increased when $\beta_1 = \beta_2$ or $\beta_1 < \beta_2$, the empirical power of t_{z_4w} decreased. In situations where $\beta_1 > \beta_2$, the empirical power of t_{z_5w} increased but t_{z_4w} reamined constant. The results of the investigation of the two-tailed tests were not in complete agreement with their corresponding one-tailed tests. A possible explanation is that the distribution of the test statistics may be skewed.

The results of this investigation are limited to the parameter values chosen. In other words, if some parameter values were changed such as $\rho_{y_oy_o}$

and the relative magnitude of β_1 and β_2 some of the results would be different. For example, the satisfactory results based on using t_{z_2w} were functions of the chosen parameter values. If the chosen values of $\rho_{y_oy_o}$, β_1 and β_2 had been .9, .3 and .9 instead of .8, .3 and .7 when $\beta_1 < \beta_2$ the values of $\rho_{\bar{z}_2w}$ would be changed from .003 to .06. The test statistic, t_{z_2w}, may be too conservative instead of valid for these new parameters.

The results of this study indicate that procedures used by process-product researchers in forming residual gain scores typically provide misleading results. Sometimes the test statistics used are too liberal and other times they are too conservative. Therefore, it is recommended that process-product researchers who wish to test for no teacher behavior effect use t_{z_4w} which involves setting $K = \hat{\beta}_1$. In addition to yielding valid Type I error rates across all conditions investigated, the procedure had reasonable power (though not as good as if K were a known constant).

BIBLIOGRAPHY

BIBLIOGRAPHY

Brophy, J. E., & Evertson, C. M. Process-product correlations in the Texas teacher effectiveness study; final report. Report No. 74-4, University of Texas at Austin, 1974.

Bryk, A. S., & Weisberg, H. Use of the non-equivalent control group design when subjects are growing. Psychological Bulletin, 1977, 84, 950-962.

Creemers, B. Evaluation of styles of teaching initial reading: An investigation into the relationship between teachers' use of a specific method and pupil achievement (with summary in English). Utrecht: Drukkerij Elinkwijk B. V., 1974.

Creemers, B. & Weeda, P. Evaluation of teaching styles: An investigation into the relationship between pupil achievement and teachers' use of a method for teaching initial reading. Unpublished manuscript, State University of Utrecht, Netherlands, 1974.

Cronbach, L. J. Research on classrooms and scholars: Formulation of questions, design and analysis. An occasional paper of the Stanford Evaluation Consortium, Stanford University, 1976.

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

Doyle, W. Classroom tasks and students' ability. In P. L. Peterson & H. J. Walberg (Eds.), Research on teaching. Berkley, CA: McCutchan, 1978.

Draper, N. R., & Smith, H. Applied regression analysis. New York: John Wiley & Sons, 1981.

Dunkin, M. J., & Biddle, B. J. The study of teaching. New York: Holt, Rinehart, & Winston, 1974.

Ebel, R. L. Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall, 1979.

Elashoff, J. D. Analysis of covariance: A delicate instrument. American Educational Research Journal, 1968, 6, 383-401.

Gage, R. M. Essentials of learning for instruction. Hinsdale: The Dryden Press, 1977.

Haney, W. Units of analysis issues in the evaluation of project follow through or there must be heresy in there some place (Contract No. OEC-0-74-0394). Cambridge, MA: Huron Institute, 1974.

Hays, W. Statistics for the Social Sciences. New York: Holt, Rinehart, & Winston, 1973.

Hornquist, K. Relative changes in the intelligence form 13 to 18. Scandinavian Journal of Psychology, 1968, 9, 50-82.

IMSL Library. The IMSL Libarary, Vol. 3. Houston: International mathematical and Statistical Libraries, 1982.

Kenny, D. A quasi-experimental approach to assessing treatment effects in the non-equivalent control group design. Psychological Bulletin, 1975, 82, 345-362.

Kessler, R. C. The use of change scores as criteria in longitudinal survey research. Quality and quantity, 1977, 11, 43-66.

Linn, R. L., & Slinde, J. A. The determination of the significance of change between pre- and posttesting periods. Review of Educational Research, 1977, 47, 121-150.

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

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

Lord, F. M. A paradox of interpretation of group comparisons. Psychological Bulletin, 1969, 68, 304-305.

Olejink, S. F., & Porter, A. C. Bias and mean square errors of estimators as criteria for evaluating competing analysis strategies in quasi-experiments. Journal of Educational Statistics, 1981, 6, 33-53.

Porter, A. C. Analysis strategies for some common evaluation paradigms. Paper presented at the Annual Convention of the American Educational Research Association, New Orleans, 1973

Porter, A. C. The effcts of using fallible variables in the analysis of covariance. Unpublished Ph.D. dissertation, University of Wisconsin, 1967.

Porter, A. C. How errors of measurement affect ANOVA, regression analysis, ANCOVA, and factor analysis. Paper presented at the Annual Convention of the American Educational Research Association, New York, 1971.

Porter, A. C., & Chibucos, T. R. Selecting an analysis strategy. In G. Borich (Ed.), Evaluation Educational Programs and Products. New York: Educational Technology Press, 1974.

Rosenshine, B. The stability of teacher effects upon student achievement. Review of Educational Research, 1970, 40, 647-662.

Soar, R. S. An integrative approach to classroom learning (Grant No. 5-R11 MH 020045). Philadelphia: Temple University, 1966.

Tucker, L. R., Damarin, F., & Messick, S. A base-free measure of change. Psychometrika, 1966, 31, 457-473.

Veldman, D. C., & Brophy, J. C. Measuring teacher effects on pupil achievement. Journal of Educational Psychology, 1974, 3, 319-324.