DOCUMENT RESUME

ED 103 440 95 TM 004 271

AUTHOR Marx, Thomas
TITLE Problems in Longitudinal Analysis: A First Look at Longitudinal Versus Cross Section Studies.
INSTITUTION Huron Inst., Cambridge, Mass.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
PUB DATE 30 Jun 73
CONTRACT OEC-0-72-0718
NOTE 19p.

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS Achievement Tests; Analysis of Covariance; Analysis of Variance; *Comparative Analysis; *Cross Sectional Studies; Data Analysis; Educational Research; Experimental Groups; Individual Differences; *Longitudinal Studies; *Measurement Techniques; Models; Post Testing; Predictor Variables; Pretesting; Research Problems; Statistical Analysis; *Statistical Bias; Suppressor Variables; Tests of Significance
IDENTIFIERS Follow Through

ABSTRACT

When two groups, initially dissimilar, undergo different treatments, can subsequent differences be partitioned in such a way that the difference between the two treatments is unbiased? This is the central problem of this paper, and it is confronted by the examination of two levels of information using a Follow Through Evaluation. The first information level contains, in addition to outcome variables (achievement tests), information on child characteristics and family background. The second contains all the variables of the first plus three achievement tests given at an earlier time. Although many educators believe it is important to use a pretest to adjust posttest scores, closer inspection reveals that this process typically explains only a fraction of the variation on posttest scores. Even if pretests substantially explain variations on concurrent variables, this does not warrant the conclusion that treatment differences based on posttests will be altered by additional information. It is concluded that a multifaceted approach may reveal the analysis or analyses best suited for a particular question. (Author/BJG)

AUG. 5 1974

THE HURON INSTITUTE
119 Mt. Auburn Street
Cambridge, Massachusetts 02138
617 661-9285

PROBLEMS IN LONGITUDINAL ANALYSIS

THOMAS MARX

June 30, 1973

Prepared for:

The Office of Education
Follow Through Evaluation
Contract No. OEC-0-72-0718

PROBLEMS IN LONGITUDINAL ANALYSIS:
A FIRST LOOK AT
LONGITUDINAL VERSUS CROSS SECTION STUDIES

The Problem

When two groups, initially dissimilar, undergo different treatments, can subsequent differences be partitioned in such a way that the difference between the two treatments is unbiased? That is, can the treatment difference be estimated free of all other antecedent and concurrent influences? The answer depends on how much we know about the initial dissimilarity. If all the variables that produced initial differences are known and well measured and the structure of their relationship with the outcomes measure is also known for both groups, the initial dissimilarity can be totally removed. In educational research this condition is rarely, if ever, met. When these variables are unknown or unobserved, not only the program but everything else that might have produced differences in outcomes between the groups is a possible cause.

This paper deals with two levels of information. The first contains, in addition to the outcome variables (the achievement tests), information on child characteristics and family background. The second contains all the variables of the first plus three achievement tests given at an earlier time.

The first level is an example of cross sectional data: a number of measurements are taken on a sample of people once.

Although the background variables in this case were measured prior to the post-test, we treat them as if they were gathered simultaneously since they could have been measured at the same time. The second level represents longitudinal data: a number of measurements are taken on a sample of people at time 1 and another set of measurements (on possibly different variables) is taken on the same sample at time 2. Does utilization of the longitudinal information -- in this case the three initial achievement tests -- change our conclusions about differences among groups of people on the outcome achievement tests?

Many educational researchers believe it is important to use pre-tests to adjust post-test scores. The measures available at the time of post-testing -- often socio-economic variables on the child's family along with her or his personal attributes -- typically explain only a minor fraction of the variation in post-test scores. Addition of pre-tests to the concurrent variables usually substantially increases the fraction of variation that can be explained. However, it does not follow from this that conclusions about treatment differences based upon the post-tests will be altered by the additional information. For this to happen the structure of the relationship between the expanded set of explanatory variables and the outcome variables must be changed by the introduction of the pre-test variables.

The Problem in the Follow Through Evaluation

The Follow Through evaluation, like other evaluations of large-scale educational interventions, is not a true experiment and thus does not involve random assignment at any stage. Since some child and site characteristics were measured we know that these characteristics are confounded with the models (treatment). We know the extent of confounding with the measured variables but without randomization we do not know the extent of confounding with unmeasured variables. Additionally, treatments and non-treatments were, for the most part, poorly defined. Many communities, sponsors, schools, etc. had unique features that may well have interacted with treatment or non-treatment. Also, many of the events that occur in the four years between entrance into and exit from the program are undiscoverable. Finally, almost 60% of the children in the evaluation disappear before they exit. The characteristics of the reduced sample may differ from the initial sample in ways related to outcomes.

In this situation we must use auxiliary information about students as well as family, teacher and community information to adjust estimates of outcomes. There are a number of ways of tapping this auxiliary information to produce less biased and/or more reliable estimates of outcomes. At this time, however, we want to concentrate on a narrow problem: the degree to which the addition of pretest information alters the configuration of post-test estimates.

The Study

The Sample. In this initial exploration we have utilized only a small segment of the data set collected by SRI -- a subsample of the Summer Study data. Our selection of this segment was based on convenience since a work tape had already been prepared for the Summer Study. The data have been reduced to a manageable subset of variables and reorganized so that it can be read by DATATEXT, a social science package of computer programs. This subsample consists of approximately 400 FT and NFT children in Philadelphia who entered kindergarten in fall, 1971. The tape includes test data from fall 1971 and spring 1972 and the parent interview from fall 1971. The children fall into one of three FT models and their comparison groups: Bank Street College of Education (0508), the University of Kansas (0803) and Educational Development Corporation (1103). While these three sponsors are only a subset of the FT models, they reflect the extremes on a continuum of classroom structure. Consequently, we felt that this was a sufficient sample for our purposes.

Techniques of Analysis. We wished to determine if the addition of pre-test scores would result in different inferences about effects of the three models. What we did in this study is contrast each of three MAT subtests given in spring 1972 with background information included as covariates, and with both background information and scores on each of the three WRAT subtests given in fall 1971 included as covariates.

Two separate analyses were undertaken. The first, designed to assess mean effects and interactions, was an unweighted means ANCOVA without inspection of the within cells regression coefficients. This means that we assumed that each of the six cells of the 2 by 3 design had the same relationship with a set of covariates without examining that assumption. The factors were program (Follow Through or non-Follow Through) and model (0508, 0803 or 1103). The ANCOVA adjusts each obtained mean by using the regression coefficients, the within cell means of the covariates, and the grand means of the covariates. In the ANCOVA we estimated the two main effects and their interaction under the assumption of a fixed model.

The second analysis used the general linear model and dummy coded variables to estimate effects for five different FT/NFT by site combinations. The one excluded combination was NFT in site 1103 since the inclusion of all six groups in the model would have made the data matrix linearly dependent. This second analysis is equivalent to an unbalanced one-way ANCOVA under the fixed model assumption with an exact least squares solution. The levels of the classification factors are five FT/NFT by site combinations.

Before presenting results we should discuss the variables used in the study. The three dependent variables were the Listening for Sounds, Reading, and Numbers subtests of the MAT Primer test battery. Subtest reliabilities are all between

.89 and .96 whether measured by split half reliability corrected by the Spearman-Brown formula or a modification of Kuder-Richardson formula 20.* (Low reliability would lead to an enlargement of error variance but would not bias estimates.)

The three fall WRAT tests used as covariates were in fact proper subsets of the full-length WRAT subtests. Items were deleted from the whole subsets by SRI because they were deemed too difficult or in some other way unsatisf· ory for the children tested. The three SRI WRAT subtests were the Spelling, Reading and Math. SRI computed Cronbach's coefficient alpha as a reliability measure for each of three subtests and in all cases it was around .95. This means that we do not have to worry about the bias that might result from not adjusting each of the SRI WRAT subtests by its reliability coefficient since the reliability is so high that the adjustment would add little accuracy to the analyses.

The background covariates were as follows:

<u>Age</u> -- the child's age in months on September 1, 1972.
<u>Sex</u> -- coded as 0 for boys and 1 for girls to allow for an effect estimate.
<u>Preschool experience</u> -- the number of months of preschool experience with a maximum of 36 months.
<u>Household size</u> -- the total number of persons living in the child's household.
<u>Mother's education</u> -- the approximate number of years of schooling completed by the child's mother.
<u>Household income</u> -- the approximate amount of income available to the household annually.

Both mother's education and household income are grouped variables.

---

* Metropolitan Achievement Tests Special Report, 1970 Edition (New York: Harcourt, Brace and Jovanovich, 1971).

Since the intervals were not chosen to make their variances equal, their use in ordinary least squares regressions will result in a violation of the assumption of homescedasticity about the regression surface and often a substantial and artificial increase in R-square.* This violation will, in our opinion, not result in enough bias to be worth worrying about, especially since adjusting for the problem would substantially increase the standard error of estimate on these variables.

T 'ast variables are the design variables. Their crossbreak determines the cell sample sizes and are set out below.

Table 1

FT/NFT by Model Numbers
Cell Sample Sizes

	Model Number 0508	0803	1103	
FT	$n_{11} = 71$	$n_{12} = 85$	$n_{13} = 82$	$n_{1+} = 238$
NFT	$n_{21} = 32$	$n_{22} = 43$	$n_{23} = 38$	$n_{2+} = 113$
	$n_{+1} = 103$	$n_{+2} = 128$	$n_{+3} = 120$	$n = 351$

The total sample of 3̄1 is that number of children who had valid

---

* J. Johnston, Econometric Methods, 2nd Ed. (New York: McGraw-Hill, 1971), pp. 228-238.

scores on all six of the subtests, the three WRATs and the three MATs. The original eligible pool was composed of 481 children.

Table 2 presents the effects of the ANCOVA. Row and column figures represent the main effects and cell figures represent the interaction effects. For each row, column and cell there are two entries: the upper one gives the adjusted effect after the background covariates have been included in the equation and the bottom entry gives the adjusted effect after both background and WRAT pre-test covariates have been included. The row effects reflect the overall differences between Follow Through and non-Follow Through for these models. The column effects are the model effects -- a misnomer since only the FT groups have the model. They reflect differences among the FT and paired comparisons for the three models, and do not involve a comparison between FT and NFT. The interactions in the FT row are the effects of each of the three models, which can be compared to each other and to the corresponding NFT cells. These are the effects we are most interested in since they form the basis for inferences about differential model effects.

Thus, instead of trying to interpret the entire table we have summarized the relevant data. Table 3 presents the adjusted means for the six cells under both analyses. The first two rows under each subtest show the results without the WRAT pretest as a covariate. The second two rows reflect the

TABLE 2

MAIN EFFECTS AND INTERACTIONS
FOR TWO ANCOVAs ON EACH OF
THREE MAT SUBTESTS*

		Model Number 0508	0803	1103	Program Effects	Grand Means (Unadjusted)
		MAT SOUNDS				
	FT	1.283	.270	-1.553	-.356	
		.881	.185	-1.067	-.435	
	NFT	-1.283	-.270	1.553	.356	
		-.881	-.185	1.067	.435	
Model Effects		.444	.741	-1.185		16.698
		.108	1.011	-1.119		
		MAT READING				
	FT	.720	.278	-.998	-.359	
		.412	.171	-.582	-.498	
	NFT	-.720	-.278	.998	.359	
		-.412	-.171	.582	.498	
Model Effects		-.068	1.084	-1.016		14.306
		-.377	1.290	-.914		
		MAT NUMBERS				
	FT	-.493	1.176	-.683	.508	
		-.787	.978	-.191	.315	
	NFT	.493	-1.176	.683	-.508	
		.787	-.978	.191	-.315	
Model Effects		-.063	1.742	-1.679		-11.466
		-.445	2.013	-1.568		

* Upper entries represent background covariates alone; lower entries represent background covariates and WRAT covariates.

results with the WRAT used as a covariate.

TABLE 3

ADJUSTED CELL MEANS

Model		05	08	11
			Sounds	
no WRAT	FT	18.069	17.354	13.604
	NFT	16.214	17.524	17.422
with WRAT	FT	17.252	17.459	14.077
	NFT	16.359	17.959	17.081
			Reading	
no WRAT	FT	14.599	15.308	11.932
	NFT	13.877	15.471	14.647
with WRAT	FT	13.842	15.269	12.312
	NFT	14.016	15.924	14.473
			Numbers	
no WRAT	FT	11.418	14.892	9.612
	NFT	11.387	11.525	9.962
with WRAT	FT	10.549	14.773	10.022
	NFT	11.493	12.186	9.773

Since we are interested in the effects of the FT models, we then took the difference between the FT and NFT adjusted means for each model on each subtest. Table 4 presents the FT/NFT differences for the two analyses.

TABLE 4

DIFFERENCES IN ADJUSTED MEANS
(FT-NFT) WITHIN EACH MODEL

Analysis	05	08	11
		Sounds	
(a)	1.855	-.170	-3.818
(b)	.893	-.500	-3.004
		Reading	
(a)	.722	-.163	-2.715
(b)	-.174	-.655	-2.161
		Numbers	
(a)	.031	3.367	-.350
(b)	-.944	2.587	.249

(a): analysis without the WRAT pretest

(b): analysis with the WRAT pretest as a covariate

This table now allows a comparison among the FT/NFT differences across models for the two analyses. For both Sounds and Reading, inferences about the relative standing of the models are the same under both analyses. The WRAT covariate narrows the separation but doesn't change the order. Thus, inferences about the size of the differences are affected. For the Numbers subtests, inferences about the relative standing of the models is affected, though not dramatically. Model 08 is the highest in both cases, but 05 and 11 change order when the WRAT is introduced. Inferences on the size of the differences are also affected.

Finally, in all three subtables of Table 4, the effect estimates change by enough to make the inclusion of the three WRAT pretests seem worth the expense and effort.

In Table 5, we present the ANOVA parts of the ANCOVA. There are again double entries within cells to represent the partial and full sets of covariates.

The thing to notice about Table 5 is that the addition of the three WRAT tests to the set of covariates increases precision on all three subtests, that is, the mean square residual always decreases. Since the degrees of freedom for main effects and interaction tests are the same no matter how many covariates we have, other things equal, an increase in precision should increase F-ratios for effects and interactions and lower the associated significance levels. And in eight of twelve instances F ratios rise and significance levels fall. However, in three instances -- Sounds and Reading interactions and Numbers program effect -- F ratios fall and significance levels rise. This suggests that the amount of bias present without the introduction of these three additional covariates is so great that it more than offsets the gain in precision.

TABLE 5

ANCOVA ANALYSIS OF VARIANCE
TABLE FOR THREE MAT SUBTESTS

	D.F.	Sum of Squares	Mean Square	F-Test	Significance
			M A T S O U N D S		
Program	1	33.910	33.910	1.091	.297
	1	48.980	48.980	1.840	.176
Site	2	202.879	101.439	3.265	.040
	2	213.965	106.982	4.018	.019
Program by Site	2	410.816	205.408	6.611	.002
	2	187.879	93.939	3.529	.031
Covariates	6	502.016	83.669	2.693	.015
	9	2089.867	232.207	8.722	under .001
Residual (Error)	339	10533.180	31.071		
	336	8945.328	26.623		
Total (After Mean)	350	11655.504	33.301		
			M A T R E A D I N G		
Program	1	34.563	34.563	2.320	.129
	1	64.101	64.101	5.562	.019
Site	2	210.426	105.213	7.062	.001
	2	250.263	125.132	10.857	under .001
Program by Site	2	158.922	79.461	5.334	.006
	2	52.150	26.075	2.262	.106
Covariates	6	510.016	85.003	5.705	under .001
	9	1687.916	187.546	16.272	under .001
Residual (Error)	339	5050.547	14.898		
	336	3872.647	11.526		
Total (After Mean)	350	5962.102	17.035		
			M A T N U M B E R S		
Program	1	69.148	69.148	3.119	.079
	1	25.684	25.684	1.527	.218
Site	2	557.184	278.592	12.567	under .001
	2	634.160	317.080	18.846	under .001
Program by Site	2	206.840	103.420	4.665	.011
	2	153.922	76.961	4.574	.011
Covariates	6	795.172	132.529	5.978	under .001
	9	2657.109	295.234	17.548	under .001
Residual (Error)	339	7515.023	22.168		
	336	5653.086	16.825		
Total (After Mean)	350	9118.956	26.054		

When one also examines the large shifts that occur in some of the F ratios that rise, it is clear that the addition of the pretests yields a less disorted picture of what is going on. For this set of analyses, we conclude that the longitudinal study is superior to the cross sectional study.

In Table 6 we review the analyses using the general linear model. The top row under each subtest presents the results for the analysis without the WRAT pretests. The second row has the results with the pretests as covariates. The column headings are now the FT/NFT by model combinations and each of the five effects is tested on a single degree of freedom. The combinations are coded as a series of five dummy variables. The upper entry in each cell is the effect estimate. This estimate equals zero for the omitted sixth group. The bottom entry in each cell is the significance level of t-tests made on 339 degrees of freedom for the first analysis (without the WRAT) and 336 degrees of freedom for the second analysis.

In order to compare model effects we have summarized Table 6 in Table 7. In this table, the row headings are the same. Each cell entry now is the difference between the FT effect and the NFT effect for each model under each analysis.

The FT/NFT differences in this table are almost identical to those in Table 4. Consequently, the conclusions drawn from Table 4 on page 10 are the same for these analyses.

TABLE 6

GENERAL LINEAR MODEL EFFECT SIZES AND THEIR SIGNIFICANCE LEVELS FOR TWO ANALYSES AND THREE MAT SUBTESTS

	FT 0508	NFT 0508	FT 0803	NFT 0803	FT 1103
			MAT SOUNDS		
Background Covariates	.6477 over .500	-1.1943 .383	-.0747 over .500	.0720 over .500	-3.8204 .001
Background & WRATS Covariates	.1908 over .500	-.7046 over .500	.3933 over .500	.8682 .461	-3.0072 .005
			MAT READING		
Background Covariates	-.0650 over .500	-.7516 .427	.6444 .414	.8126 .353	-2.7367 under .001
Background & WRATS Covariates	-.6250 .385	-.4344 over .500	.8006 .259	1.4580 .061	-2.1829 .002
			MAT NUMBERS		
Background Covariates	1.4303 .146	1.4241 .218	4.9034 under .001	1.5201 .154	-.3740 over .500
Background & WRATS Covariates	.7742 .373	1.7223 .089	4.9992 under .001	2.3951 .011	.2267 over .500

TABLE 7

FT EFFECT-NFT EFFECT FOR
THE GENERAL LINEAR MODEL FOR
TWO ANALYSES AND THREE MAT SUBTESTS

	05	08	11
		SOUNDS	
(a)	1.8420	-.1467	-3.8204
(b)	.8954	-.4749	-3.0072
		READING	
(a)	.6866	-.1682	-2.7367
(b)	-.1906	-.6574	-2.1829
		NUMBERS	
(a)	.0062	3.3833	-.3740
(b)	-.9481	2.6041	.2267

(a) analysis without WRAT pretest

(b) analysis with WRAT pretest as a covariate

Recommendations

The study reported here is just a start. Much more work using different strategies and samples needs to be done before researchers begin to get a feel for the magnitude of the data collection effort they must make for conclusions to be given credence.

Just confining ourselves to the data from the Follow Through Evaluation, there are many ways to sample individuals, other units of analysis or variables from the data tape and many ways to subject the different samples of units and variables to quantitative analyses. At this point a multi-faceted approach may reveal the analysis or analyses best suited for a particular question.