DOCUMENT RESUME

ED 202 079 EA 013 391

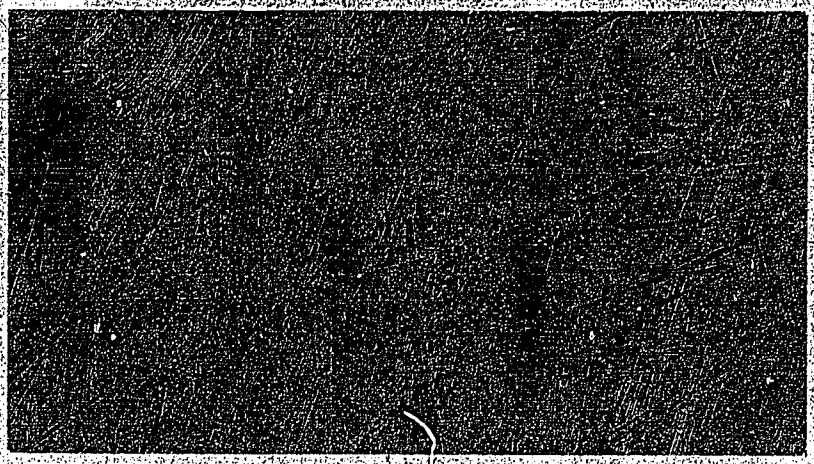
AUTHOR Jovick, Thomas D.
TITLE Longitudinal Data Analysis: Approaches to Data Analysis in Project MITT.
INSTITUTION Oregon Univ., Eugene. Center for Educational Policy and Management.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE 78
NOTE 46p.; For related documents, see ED 172 425 and EA 013 390.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Analysis of Covariance; Analysis of Variance; Elementary Education; *Longitudinal Studies; *Multiple Regression Analysis; Path Analysis; Research Problems; School Organization; *Statistical Analysis; Team Teaching
IDENTIFIERS Management Implications of Team Teaching Project

ABSTRACT

This report explains why the Management Implications of Team Teaching (MITT) project chose multiple linear regression and path analysis to analyze through-time relationships among variables, and why it rejected repeated-measures analysis of variance (ANOVA) and difference scores over time. Project MITT examined governance and work structures for five time periods from 1974 to 1976 in 29 elementary schools, 16 of which had introduced team-teaching (or unitized) methods in 1974. To analyze longitudinal changes among variables and schools, the project's statistical techniques had to take account of small sample size and multiple time periods; they also had to control for pre-1974 differences among the schools, changes in variables because of unitization, and differences in variable means and ranges. All of these factors interfered with comparisons of unitized and nonunitized schools and distorted relationships among the variables. Hierarchical multiple linear regression solved these problems by relating variables to one another both over time and in order of explanatory power. Path analysis using lagged multiple linear regression helped to test postulated relationships through time and explore for further relationships. Four appendices discuss ANOVA, difference scores, path analysis, and corrections used for data cyclicity. (Author/RW)

project MITT



Center for Educational Policy & Management

UNIVERSITY OF OREGON / 1472 KINCAID, EUGENE, OREGON 97401

LONGITUDINAL DATA ANALYSIS: APPROACHES TO DATA ANALYSIS IN PROJECT MITT

By

Thomas D. Cook

Spring, 1978

Center for Educational Policy and Management
University of Oregon
Eugene, Oregon

The research reported herein was supported in part by funds from the National Institute of Education, U. S. Department of Health, Education, and Welfare. Opinions expressed in this report do not necessarily represent the policies or positions of the National Institute of Education, nor should any official endorsement of the report be inferred from the reporting agencies.

The University of Ore n, a member of the [illegible]on State System of Highe Edu [illegible]on, prohibits d imination based on race, c [illegible]or, religion, sex, age ndicap, or national origin. s p[illegible]icy implements [illegible]ous fed al and state law nd executive orders [illegible]cluding Tit IX and its regu[illegible]tions) and applies to ployment, admission, education, nd facilities. Direct nquiries to Myra T. Willard, D ector Affirmative Action Office, Oregon Hall, Universi of Oregon, Eugene, Oregon 97403 Telephone: (503) 686-3123.

TABLE OF CONTENTS

Introduction

This report originated in the search for an appropriate strategy for analyzing through-time relationships among selected variables in the MITT (Management Implications of Team Teaching) study. It explains the rationale for our use of multiple linear regression and, at times, path analysis as the means of sorting out through-time relationships and discusses some of the less fruitful approaches we had considered at first.

MITT had collected data concerning the governance and work structure in 29 elementary schools, 16 of which implemented a multiunit form of organization among the teaching staff in the fall of 1974. To strengthen our confidence in inferences about the effects of adopting the multiunit organization, we collected data in the spring of 1974, when units had not yet been formed, and every six months thereafter for two years. Thirteen of the schools adopted no such innovative structural change over the length of the study and served as controls matched by district to the experimentals whenever possible (Packard et. al., 1976).

Because of the difficulties in getting a through-time individual-level file together, the majority of MITT's early analyses used only school level indices, although some variables existed only at the school level, e.g. extent of Collegial Decision Making, others had to be aggregated as means, e.g. extent of Classroom-related Communication. This immediately put a constraint on the effective sample size for any analytical strategy we planned to use.

More substantive reasons existed for employing a school level analysis. We had conceptualized some major variables of interest as properties of the organization and expected they would change over time in response to the anticipated school-wide installation of a multiunit structural organization among the faculty. Although the changes we were investigating relied upon activities of individual teachers, the implementation was to be school-wide, reflecting the behavior of most, if not all, teachers in the school.

Furthermore, the schools were units of analysis which remained throughout the course of the study, even though the teacher turnover gave us a slightly different staff composition at each wave. In fact, only about two-thirds of the faculty in all unitized and conventional schools at T1 were present in the schools at T5. By using the school as the unit of analysis we did not have to confine our information about variables to that given by this two-thirds faculty cohort.

This does not imply we eliminated the individual teacher as a unit of analysis; many of our cross-sectional and longitudinal analyses used the teacher as the unit (Packard et. al., 1976; Packard et. al., 1978). This was especially true for the teacher attribute and perceptual variables which conceptually characterize properties of individuals rather than organizations. For the school-level analyses, we aggregated many of these to depict mean levels of selected teacher attributes in each school. (Packard et. al., 1976; 1978.)

Certainly, the linear regression strategy which we adopted for longitudinal data analysis applied equally to the individual level, but because of the larger sample for analysis, problems in restricting the number of independent variables did not apply. At the time we were selecting a longitudinal strategy, however, the individual through-time file did not even exist and hence was not a major focus of our concern. Furthermore, some of the problems and alternative strategies we encountered transcend the unit of analysis as a consideration.

Selecting the Method of Longitudinal Analysis

Our preliminary queries about across-time analyses initially centered upon the detection of experimental-control differences at the various points in time while taking into account temporal differences in variation in the dependent variable of interest. Our naivete led us to attempt to fit the Reapeated-Measures ANOVA to our design but that model was eventually deemed wholly inappropriate (See Appendix A).

Two other strategies struck us as viable options for analyzing differences between experimentals and controls and for relating change in one or more variables to change in another. One was the use of difference scores, created by subtracting scores at one point from those at an earlier point, which would be used in some type of correlational analysis or group comparison. This approach also proved unacceptable (See Appendix B).

We settled upon the generalized multiple linear regression. The approach relates the status of a dependent variable at one point in time to the status of the same and/or other variables at previous points in time. Its use in analysis of covariance is quite amenable to the study of gain or loss (change) as a function of treatment. (Pelz and Lew, 1970.) In addition it permits the assessment of curvilinear and contingent/interactive relationships (Cohen and Cohen, 1975; Amick and Walberg, 1975).

The analytical regression strategy we selected is called hierarchical regression analysis. A dependent variable is regressed on several independent variables in a particular order. For example, if Y were regressed on X1, X2, and X3, each in that order, a hierarchical analysis will provide essentially three types of information.

One is the total proportion of variance in Y that is accounted for by the three variables together, R^2. Another is the increment in the proportion of variance explained due to the addition of a variable. This means one essentially has three regression equations: Y with X1, Y with X1 and X2, and Y with X1, X2, and X3. The increments are the proportion of variance explained by X1 alone, that added by X2 after X1 is already entered, and that added by X3 after X1 and X2 are already in the equation.

A final important source of information are the regression coefficients, indices of the unique "effect" of each independent variable upon the dependent variable. The term "unique" indicates a coefficient that reflects the directional relationship of an X on Y controlling for the amount of variance in Y that the X's share with each other. The fact that "sharing" occurs is partly reflected in the fact the the independent

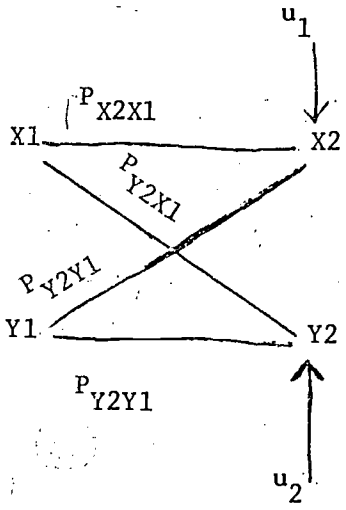
variables are usually correlated. The coefficients are called b-weights when variables are expressed in raw score form and beta weights when all variables have been converted to standardized scores with means equal zero and standard deviations equal one.

Path analysis, a method for testing hypothesized causal relationships with the use of the multiple regression, held some promise for aiding our assessment of longitudinal relationships. In some instances we employed this approach to test a set of carefully postulated relationships, in others we searched for relationships in a more exploratory fashion (See Appendix C).

Lagged Multiple Linear Regression Analysis

Heise (1970) presented a model for using path analysis to assess through-time causal relationships when one has two waves of data. Pelz and Lew (1970) extended his model to cover multiple waves of data.

Heise dealt with a two-wave two-variable system as diagrammed below -- the subscripts "1" and "2" indicate earlier and later points in time respectively.



His approach is actually an extension of path analysis to longitudinal data by means of lagged regressions. The regression estimates are obtained through a set of multiple regression analyses -- X2 is regressed on X1 and Y1 in that order, Y2 is regressed on Y1 and X1 in that order. The standardized regression coefficients or betas resulting from the analyses are estimates of the path coefficients and are represented as p's; as with the typical cross-sectional path analysis, his strategy uses beta weights as the meaningful coefficients.

Path p_{X2Y1} represents the impact of variation in Y at Time 1 on variation in X an Time 2; p_{Y2X1} represents the impact of variation in X at Time 1 on variation in Y at Time 2. One can compare these empirical betas or "path coefficients" and infer direction and magnitude of influence, which may involve only X1 to Y2, only Y1 to X2, both, or neither.

The paths p_{X2X1} and p_{Y2Y1} represent the temporal stability in X and Y respectively. Large positive stability coefficients suggest not much happened during the interval to disturb the original distributions; that is, variations at T1 for the most part determine variations at T2. Low stability coefficients would suggest that the distributions for each variable changed considerably between measurements.

One of the assumptions of the model is that the measurement lag between Tn and Tn+1 matches closely the actual causal/relationship lag. Pelz and Lew (1970) expanded on this assumption with a Monte Carlo study of the model employing several waves of data and lags of different lengths.

They defined a population in which X1 caused Y2 but Y1 did not cause X2 and specified a causal interval of four units; they also specified population betas for the relationships of X1 to X2 and X1 to Y2. This enabled them to investigate the sizes of betas one would obtain if he were studying a causal lag that were either shorter or longer than the actual one. In addition, they also obtained betas for the relationship of Y1 on X2 to see what evidence the empirically discrepant lags might produce about the existence of this relationship when in fact it did not exist.

Equilibrium

As a guide for what to expect given certain levels of long-term and short-term stabilities, inaccurate measurement lag, and true population causal relationships, the Pelz and Lew study initially showed promise of offering us some utility. However, because it addressed a system of relationships in equilibrium, uninterrupted by contrived/planned change or trauma (as opposed to emergent change) we grew increasingly doubtful of the interface with the processes we were examining.

---

*This last observation we infer after reading both Heise's and Pelz and Lew's discussion of the model.

Under a system of equilibrium, the values of variables in schools undisturbed by a major restructuring would fluctuate through time around some level. In a system altered by some innovation, values of variables may increase or decrease over time eventually to settle into another condition of equilibrium around some new higher or lower (or perhaps the original) level.

Contingent rising and falling of variables as a function of each other would also describe a state of equilibrium in the schools. If relationships exist across waves between a pair of variables, one would expect to find increases in one variable followed later by increases in another (were the relationship positive) or by decreases (inverse relationship) in another. In unitized schools we expected the introduction of the innovation to disrupt the equilibrium, upsetting normal fluctuations in, and normal contingencies among, the variables; after a time, the variation and relationships would settle back into another state of equilibrium.

The change introduced by the innovation studied by MITT did not occur at one point in time; because the units continued to exist beyond the point of their formal establishment in the school, any new equilibrium level would evolve in their presence. Regardless of the level around which values fluctuated, the new equilibrium in the unitized schools would represent the status under a qualitatively different situation than in the nonunitized schools.

Although the Pelz and Lew formulation helped us determine the strategy and clarify some assumptions of our longitudinal analyses, the equilibrium aspect left us in doubt as to their models applicability as a guide for our analyses. Conceivably, only our control schools could be considered in a general state of equilibrium, particularly during the

first year. A risk was that the basic Heise model which appeared suitable to a system in equilibrium was unsuitable for a system disrupted, in part, by planned school-wide change.

Coleman (1968) developed a mathematical treatment for analyzing change which drew heavily on the use of multiple linear regression output in equations from calculus. A basic notion in the formulation of his approach was the idea of systems in equilibrium. We attempted some preliminary analyses using one of the models he discussed and found some of those results consonant with expectations he laid out. Nevertheless, we had reservations about the applicability of his formulation, the particular model of his we selected to use, and the proper interpretation of the results; moreover, the complicated presentation made us doubt our own understanding of many of his formulations.

Stability

Pelz and Lew also discussed considerations of short and long-term stability in a variable which are reflected in the autocorrelation between different waves -- adjacent-wave correlations reflect short-term stability, longer discrepancies reflect long term stability. If the autocorrelation drops to zero as the time lag increases, then long-term stability is low or does not exist; if it drops to some constant value, then it exists to some degree depending upon the size of the correlation. They interpret long-term stability in terms of persistent characteristics of individuals such as personality and I.Q. The analogy to schools might be something like school climate or control structure or more pervasive immutable characteristics such as district wealth, school size, staff characteristics.

Some variables which characterize the school may change on a cyclical nature. For example, in our design, we would expect to find greater teacher turnover between than within years. Similarly, we would expect many of the decisions about year-round routines to be made in the fall. In combination with mean trends, autocorrelations could be used to assess the cyclical nature of school characteristics.

A variety of patterns could appear. The successive rise and fall of the mean level of a variable accompanied by a high wave-to-wave autocorrelation would signal the presence of a cycle typifying most of the schools. Crests in the fall of the year would be followed by troughs in the spring, troughs in the fall would be followed by crests in the spring. High correlations between seasons but not between adjacent waves would also suggest a cyclical pattern.

Weak autocorrelations signal that the differences between means through time do not necessarily reflect what actually goes on in each school. For example, if the overall means for a variable stay the same between two waves but the autocorrelation is weak, then we can infer that the scores for all schools do not tend to remain the same; if they did we would expect a high correlation.

Use of Lagged Multiple Linear Regression on MITT Data

MITT used regression analyses to address two general types of goals: one was the detection of T5 differences between unitized and nonunitized schools, the other was the assessment of lagged relationships among

variables. The MITT Project had data that could be analyzed at different levels; those of primary interest were the individual, the unit (in which case the analysis naturally was confined to the unitized schools), and the school. At the school level, the sample size, 29 at best, constrained the number of independent variables we could use. We usually limited the number to two and, in the case of regressions using the autoregression term, this meant we generally had one independent variable of central interest.

Controlling for Pre-Unitization Differences

The availability of T1 data provided additional information about variance in the dependent variable at some later time, Tn, and afforded us the opportunity for a more powerful analysis. Because the two groups of schools differed at T1 on several variables, we could not be sure that any differences detected at T5 could be attributed to effects of unitization. The statistical determination of unitized-nonunitized differences had to take the pre-unitization differences between the two sets of schools into account. To do this we employed the hierarchical multiple regression approach by regressing T5 values of a dependent variable first on the T1 values of the same variable and then on a second variable, a dummy-coded vector with 1's for unitized and 0's for nonunitized schools. This was our unit organization variable or, as we called it, EXPCON.

Since the procedure is the regression application of the analysis of covariance, our interest was in the increment in the proportion of variance accounted for by EXPCON beyond that explained by the T1-T5 autocorrelation (Kerlinger and Pedhazur, 1973). In the same regression equation we could check for an interaction effect to determine if the influence of unit organization on T5 values were contingent upon T1 values. The absence of an interaction is necessary for use of the analysis of covariance; had we found significant interaction we would do other further analyses to assess its nature since the presence of any main effects would have been uninterpretable.

Lagged Relationships Among Variables

We also sought to assess lagged relationships among a variety of variables in the study. For this purpose we usually assumed that the most relevant through-time influence on a dependent variable came from variation in the immediately previous wave. Our approach examined adjacent-wave contingences among the selected variables. The regression analysis attempted to find evidence that the Tn+1 variation in a dependent variable was influenced by variation in other variables that occurred at Tn.

The Tn-Tn+1 autocorrelation reflected the extent to which the level of a variable at a particular wave came about in response to or at least as some predictable function of its level at the immediately previous wave. A low autocorrelation would suggest that the level of a variable in a school at Tn+1 is a function of something other than its level at Tn; a high autocorrelation would suggest the level of a variable came

about at least in partial response to its previous level and perhaps in response also to some other variables.

There existed two situations for which we generally chose not to include the autoregression term. One was the case in which we formally applied a path analytic procedure to test a model of postulated lagged relationships which specifically excluded the autoregression; including it in such instances would have altered the model under examination. The other was the case in which the independent variable(s) of central interest at T_n correlated strongly with the autoregression variable. Under such circumstances of simultaneous variation we would be unable to separate out the effects of the key independent variable; the betas for each essentially would be uninterpretable because they would be showing only effects of each controlling for shared variation with the other, and the amount of shared variation controlled would tend to be large.

Finally, the Companion Study* of the MITT Project carried out several regressions to determine predictors of success in teaming in the 15 unitized schools. The predictor variables were formulated to characterize schools and hence limited the analysis to the school level; furthermore, the focus of the Companion Study on the unitized schools limited the sample size to 16 schools at best.**

---

*See Packard et. al. (1976) for a more detailed description of the Companion Study.

**Two of the original 16 unitized schools discontinued their unit structure in the second year; another provided us no data concerning instructional interdependence.

Our analytical approach assumed that whatever lagged effects or autocorrelations we observed were characteristic of all schools at the same time. Thus, we expected variables themselves to change at about the same time in all schools and contingencies among variables to arise and disappear at the same time in all schools. An alternative formulation, however, is that the schools were out of synchrony in the changes that occurred in each. Over a particular lag, a large change in a variable in some schools may have been absent or in the opposite direction in other schools; a contingency among two variables over a particular lag in some schools may not have appeared until some later lag in other schools. This alternative perspective was pursued as part of the Companion Study (Packard et. al., 1978, Ch. 8).

Combining Unitized and Nonunitized Schools

Partial Confounding Due to Unitized-Nonunitized Differences

The apparent impact of unitization in changing some of our major variables like the number of pairs of Instructionally Interdependent teachers (NPI) and the percent of Collegially-made decisions (COLL) posed potential analysis problems. Both of these variables showed a change in the experimental schools which lasted for the duration of the study following the installation of the units. In the second year, Collegiality increased slightly again and Interdependence decreased slightly in the unitized schools but both remained significantly above the levels found nonunitized schools.

When using all 29 schools together, however, a problem of ambiquity in relationships existed among some main school-level variables of interest. This can best be illustrated with the Collegiality (COLSUM) and Interdependence (NPI) variables. The fact that unitized schools are higher than nonunitized schools on both variables distorts the relationship between the two when examined across all 29 schools.

Across all schools, COLL and NPI were positively correlated within any wave; but these correlations are spurious, reflecting more the similarity in level on the two variables within experimentals and controls than any actual relationship between them. High correlations of both COLL and NPI with the unitized-nonunitized classification (EXPCON) confounds the observed relationship between each of them; their variances partly reflect the wide differences in mean levels between the two types of schools.

This contamination with EXPCON differences also frustrates the assessment of through-time stability of a variable because the autocorrelations reflect more the stability of the set of unitized schools being at a high level and the set of nonunitized schools being at a low level on the variable. The stability coefficient reflects more the enduring categorization of schools as unitized or nonunitized through time.

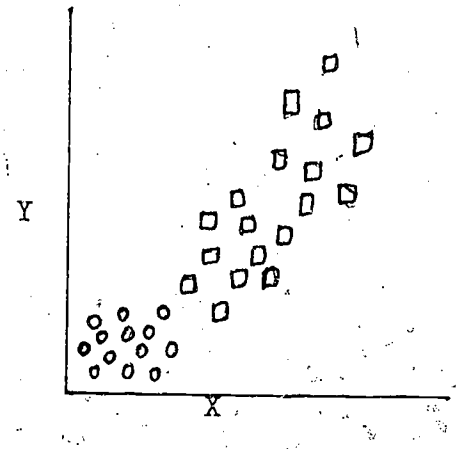
These differences suggested the possibility that the wave-to-wave relationships among the variables changed in the experimentals. Conceivably changes could occur both in the stabilities of the variables and in the cross-wave influence between any two different variables. The

usual approach for assessing this is to test the interaction between the EXPCON variable and a covariate in their relationship to the dependent variable. A significant interaction implies that the process under study differs for experimentals and controls and, therefore, that the two sets of schools remain separate for analysis purposes.

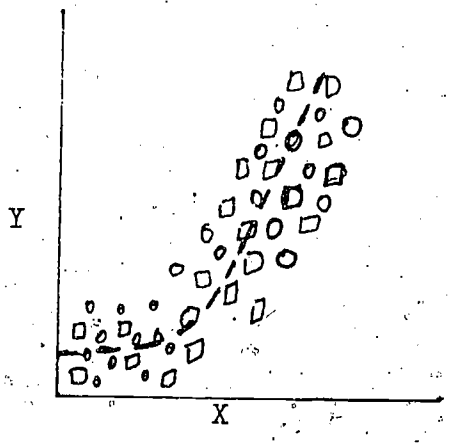
Another problem occurred which had implications for such an interaction analysis. For variables like Collegiality and Interdependence, the nonunitized schools had both lower mean values and more restricted ranges than the unitized schools on each. In this type of circumstance, any differences between experimentals and controls in correlations between variables or in stabilities (autocorrelations) for each variable may actually be a function of the different ranges and means for the two types of schools. Indeed, what might look like a strong interaction may actually reflect some curvilinear relationship which goes undetected because the range covered in one type of school essentially starts where the other leaves off.

Figure 1-a depicts a possible case. The circles represent the control schools, boxes represent the experimentals. If a covariance analysis were run on these data it would show a significant interaction -- the relationship between X and Y would be about zero for controls but positive for experimentals. However, had we a sufficient range in both experimentals and controls on X, the relationship between X and Y may actually prove to be curvilinear (Figure 1-b) or linear (Figure 1-c) for both types of schools or, indeed, different in each (Figure 1-d). When the observed data appears as in Figure 1-a, there is no way of statistically sorting out the true relationship.

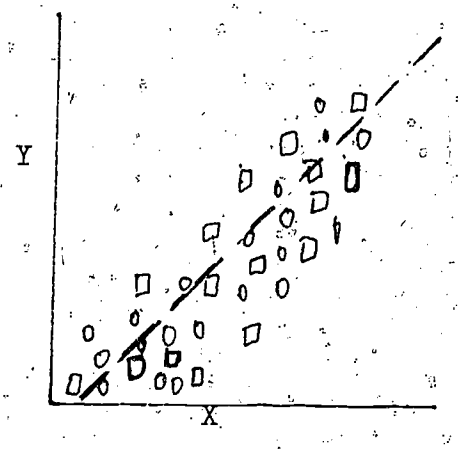
Figure 1: Relationships Between Variables in Experimentals and Controls: a) actual data with restricted ranges, b-d) possible relationships with fuller range of variation on X in both experimentals and controls.



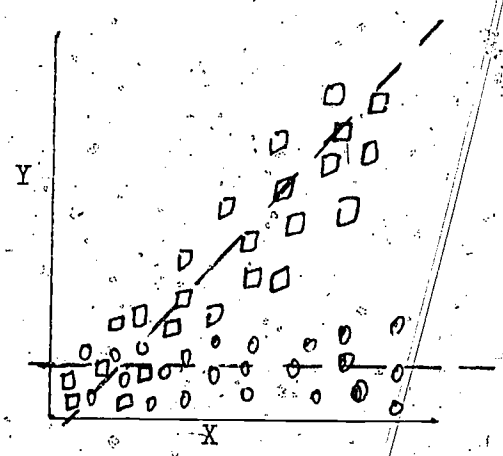
a) observed



b) possible curvilinear



c) possible linear



d) possible interaction

Where the range problem occurred, an interaction analysis was therefore deemed inappropriate since no confidence could be placed in any significant interactions found. In order to have a substantial range of variation and thereby lend somewhat more confidence to the analysis, the schools were kept combined for a preliminary examination thus assuming no interaction existed.

The point of concern, however, is still that differences in variables between unitized and nonunitized schools potentially distort the observed relationships among the variables themselves. In such cases, Cohen and Cohen (1975) would contend we cannot place much confidence in the observed relationships among the two variables or in the stability coefficients of either when the two sets of schools are combined. Any analysis should attempt to remove this source of distortion before assessing the relationships among the variables or their stabilities.

Corrections for Confounding

Decisions therefore had to be made on how to remove the distorting influence of EXPCON and whether to remove it from both independent and dependent variables. The simplest solution would be to remove it only from the independent variables. This course of action would require only regression of the dependent variable at Tn onto both the autoregression and the independent variable at Tn-1. The nature of the regression procedure would give the unique influence of each on the dependent variable

controlling for the amount of variance in the dependent variable they share together. Since this shared variance is primarily a function of the mutual correlation of all variables with EXPCON, we would be assured the analysis controlled for the most part for its distorting influence.

A more rigid form of control would be to include EXPCON as another variable in the regression equation but this would increase the number of variables probably needlessly. This strategy, however, would have to be used if we were interested in examining the relationship between say collegial decision making or interdependence and some other variable uncorrelated with EXPCON and still wished to control each for EXPCON. This also would introduce the possibility that we would control for some of the effect in which we were interested however.

However, for a path analysis, these regression strategies appear inappropriate. In a causal model what is in one place a dependent variable can become an independent variable in another. Performing the straightforward regressions for the analysis would, in effect, keep such a variable contaminated/distorted with the experimental-control differences in level when it is a dependent variable but residualize it as an independent variable. A more appropriate strategy would be to first residualize all variables by EXPCON for waves in which they show the strong experimental-control differences.

For purposes of a cleaner interpretation of any single regression equation, Cohen and Cohen (1975) contend this is the correct strategy. Since the difference in levels for experimentals and controls

confounds relationships we seek, this difference should be removed from both independent and dependent variables. This is accomplished most easily by residualizing on the EXPCON variable at each wave. Cohen and Cohen call this and Analysis of Partial Variance (APV) since a portion of variance affected by a confounding variable is removed from all contaminated variables and the relationships are assessed on the basis of the remaining portion of variance in each.

REFERENCES

Amick, D.J. and Walberg, H.J. Introductory Multivariate Analysis for Educational, Psychological, and Social Research. Berkeley, CA.: McCutchan Publishing Co., 1975.

Cohen, J. and Cohen, P. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. New York: John Wiley & Sons, 1975.

Coleman, J.S. "The Mathematical Study of Change." Methodology in Social Research by H.M. Blalock and A.B. Blalock (eds.). New York: McGraw-Hill, 1968.

Cronbach, L.J. and Furby, L. "How We Should Measure 'Change' -- or Should We?" Psychology Bulletin 74(1), 68-80, 1970. (Also Erratta, Psychology Bulletin 74,218, 1970.)

Hannan, M.T. Aggregation and Disaggregation in Sociology. 1971.

Hannan, M.T. and Young, A.A. Estimation in Panel Models: Results on Pooling Cross Sections in Time Series. Technical Report #51. Stanford University, n.d.

Heise, D.R. "Causan Inference from Panel Data." Sociological Methodology 1970 by E.F. Borgatta and G.W. Bohrnstedt (eds.). San Francisco: Jossey-Bass, 1970.

Jovick, Thomas. "Creating Indices from the Control Structure Interview through Data Collapsing and Multidimensional Scaling: Approaches to Data Analysis in Project MITT." Research report for Project MITT. Eugene,OR.: Center for Educational Policy and Management, University of Oregon, 1978.

Kerlinger, F.N. and Pedhazur, E.J. Multiple Regression in Behavioral Research. New York: Holt, Rinehart and Winston, Inc., 1973.

Nunnally, J.C. Psychometric Theory. New York: McGraw-Hill, 1967.

O'Connor, E.F., Jr., Response to Cronbach and Furby's "How We Should Measure 'Change' -- or Should We?" Psychology Bulletin 78(2), 159-160, 1972.

_____. "Extending Classical Test Theory to the Measurement of Change." Review of Educational Research 42, 73-97, 1972.

Packard, J.S.; Carlson, R.O.; Charters, W.W., Jr.; Moser, R.H.; and Schmuck, P.A. Governance and Task Interdependence in Schools: First Report of a Longitudinal Study. Eugene, OR.: Center for Educational Policy and Management, University of Oregon, 1976.

Packard, J.S.; Charters, W.W., Jr.; and Carlson, R.O. The Control of Teaching: A Study of Innovating Elementary Schools. Eugene, OR.: Center for Educational Policy and Management, University of Oregon, 1978.

Packard, J.S.; Charters, W.W., Jr.; and Carlson, R.O. Management Implications of Team Teaching: Final Report. Eugene, OR.: Center for Educational Policy and Management, University of Oregon, 1978.

Pelz, D.C. and Lew, R.A. "Heise's Causal Model Applied," in Sociological Methodology 1970 by E.F. Borgatta and G.W. Bohrnstedt (eds.) San Francisco, CA.: Jossey-Bass, 1970.

Wiley, D.E. and Harnischfeger, A. Post Hoc Ergo Proptor Hoc: Problems in the Attribution of Change. Report #7, Studies of Educative Processes. CEMREL Inc., 1973.

_____and Hornik, R. Measurement Error and the Analysis of Panel Data. Report #7, Studies in Educative Processes. CEMREL Inc., 1973.

APPENDIX A:

Repeated Measures ANOVA and the MITT Data

Historically, repeated measures designs were developed to examine the same unit of observation under several different treatments, conditions, or trials. For example, several different ways of learning a certain type of material or several trials of rehearsal on a list of words comprised the repeated treatment for each subject. In these cases, the researcher must randomize the order of presentation of the treatments and of the words in the list. The design usually characterizes experiments concerned with accounting for effects due to learning, transfer, and fatigue. A main effect for the repeated dimension indicates that, regardless of the order in which the subjects received the treatments or conditions, their scores consistently increased or decreased. An increase is typically interpreted in terms of rehearsal/practice/transfer of learning concepts; a decrease is typically interpreted in terms of fatigue/motivational concepts.

The pre-post design common to educational field research and evaluation does not really fit this paradigm. For one thing, the levels of the repeated dimension do not coincide with the administration of a treatment or condition; the treatment, rather, intervenes between a pair of points in time.

If MITT were to use the repeated measures ANOVA, any main effort found for the repeated factor would pose interpretive problems due to this lack of correspondence in this aspect of the design models; although it would indicate that scores changed (increased or decreased) over time, no useful concepts like practice, transfer of learning, fatigue or motivation exist to which we could reasonably attribute that change.

We could not unambiguously ascribe the effect to unitization for approximately half the schools since the repeated dimension main effect reflects a through-time averaging over both unitized and nonunitized schools. If anything, we would expect to find such across-time effects due to the unitized-nonunitized distinction showing up in a significant interaction; however, the pre-post design for the research and the nature of the treatment in a study like MITT guarantee a time-by-treatment interaction.

Moreover, the use of the repeated measures model to analyze pre-post or any time-to-time data violates assumptions crucial to a repeated measures paradigm. One of these is that the correlations among the levels of the repeated dimension (trials, treatments, conditions) are the same. With random ordering of the trials or treatment conditions for each subject, there need be concern over the effects of one trial on another. The covariation among trials is distributed throughout the sample and, as one of the powerful characteristics of the analysis, is accounted for by partitioning it out of the within-subject variation. In the pre-post type of design, the order of the data collection waves cannot be randomized among the schools; any covariation among them will therefore be nonrandom.

The repeated measures ANOVA falls short also because one of its distinctive characteristics lies in its great sensitivity to detecting within-subject effects but relative insensitivity to detecting between-subject or, in our case, unitized-nonunitized effects. The covariation among trials represents explained variation due to variation within each subject and the computational procedure extracts it, thereby reducing the

unexplained or residual portion of variance; consequently, the analysis reduces the size of the residual term by an amount determined by the degree of covariation among the trials. This smaller residual term, as a divisor for the F-statistic used to test within-subject effects, is smaller than it would be had there been no repeated dimension and will increase the chances of finding a significantly large F value.

APPENDIX B:

Approaches to Analysis of Change

We spent some time working through the Cronbach and Furby (1970) article on the measurement of change to see what implications it had for MITT analyses. This appendix reflects our thinking on their strategy in addition to comments on related approaches found in other articles.

Our assessment was that we proceed in our longitudinal analyses without worrying about using their strategy for estimating true scores. Given the characteristics of the MITT design and variables we saw no guarantee that the Cronbach-Furby method nor any of the other methods mentioned in this Appendix would provide us with more accurate and unambiguous estimates than we were presently obtaining with multiple linear regression.

The Case Against Change Scores

According to Cronbach and Furby, change scores, residuals, and base free measures should not be used in statistical analyses. They will give either the same results as an analyses on the original data or results more difficult to interpret.

The relationship between change and initial status can be more simply expressed in terms of the relationship between initial and final status ($B_{GX}=B_{YX}$)*; the relationship between change and another variable, W, is likewise more simply expressed as the relationship between final status and that variable controlling for initial status ($B_{YW.X}$).

*Here, G = gain or change, X = initial status, Y = final status

In studying gains as consequences of treatments, one is interested in the null hypothesis that experimentals show the same effect as controls. The question, then, is whether true final status (Y_T) scores vary from group to group. Experimentals and controls can be expected to show some measure of change over time in relation to their initial status; although the final status may not be a direct consequence of initial status, in part it is predictable from it regardless of the particular "treatment" being imposed.

The analysis of covariance takes this predictable variation into account and then compares the deviations of observed scores from the prediction between the groups. This is the strategy we have taken with the MITT data. If experimentals and controls differ markedly in the deviations then the difference is attributed to the experimental-control distinction, although this in itself does not explain what it is about the distinction that actually produces those differences.

Nature of the Problem

The basic thrust of the Cronbach-Furby article is the same as that in John Meyer's recommendation* to use Michael Hannan's (197) longitudinal analysis approach and as that spelled out by Wiley and Hornik (1973) -- to use all pertinent information to get a handle on measurement error

*Personal communication

and thereby derive more accurate estimates of true scores. Their proposals essentially describe measurement models for relating true variables to their measured values and rely upon assumptions of classical test theory. The analysis problem is that measurement errors may have large distorting influences in the assessment of relationships among variables if they are not explicitly taken into account.

Each observed score is considered to be a combination of true score plus measurement error. Over all measurements there occurs a distribution of observed scores and of errors. A true score for an individual/school is thought of as the average score over a large number of repeated measurements of a variable at a particular time point.

Strategies for Estimating True Scores

Correction for Attenuation

This is the simplest and most straightforward strategy. There are two approaches.

One involves calculating the correlation between two variables that would occur if they were both perfectly reliable. It entails using the reliability coefficients in the following formula:

$$\bar{r}_{12} = \frac{r_{12}}{\sqrt{r_{11}r_{22}}} = \text{corrected } r_{12}$$

Any planned regression analyses would then employ these corrected correlations. Some controversy exists, however, over the use of such corrected correlations.

(1) One may be fooling oneself into believing that a better correlation has been uncovered than the one actually obtained. Nunnally (1967) notes that correlations corrected for attenuation seldomly differ much in magnitude from the obtained correlations.

(2) The correction itself may be poor if the reliability estimates themselves are poor. Moreover, since the possibility exists for reversals in signs for regression coefficients and partial and part correlations if one uses corrected rather than actual correlations, he must have confidence in the reliabilities in order to have confidence in the regression output from corrected correlations.

(3) Nunnally adds that in prediction type problems it may be inappropriate to correct for unreliability in the criterion since the issue is to predict or explain scores on that variable as they actually exist not as they would exist were the test perfectly reliable. He seems to be particularly addressing prediction problems related to selection decisions.

The second approach involves obtaining estimates of unbiased scores because obtained scores tend to be biased, i.e., high scores tend to be higher than their true score counterparts and low scores tend to be lower. Conceptually, unbiased scores are those that people (schools/units) would obtain if they were administered all possible tests having equal numbers of items sampled randomly from the same domain -- they are estimates of true scores. In the formula below,

$$t' = r_{XX}x$$

$x = (X - \bar{X})$ and t' is a true score estimate in deviation score units. By adding t' to $\bar{X}$ one obtains an estimated true score.

Nunnally recommends that estimating true scores is necessary only in longitudinal analyses where one is interested in contrasting the changes between groups. To correct obtained scores O'Connor (1972) advocates using the test-retest reliability; if it were unavailable then some measure of internal consistency would be satisfactory.

Cronbach and Furby Method

These authors extend the idea of correcting raw scores. Their strategy is to estimate true scores for independent and dependent variables within experimental and control groups and then to enter these true scores into regression equations. Their calculations for a true score on any variable employs more information than the instrument's reliability coefficient alone.

Let X1 and X2 represent time 1 (T1) and time 2 (T2) scores on variable X. Unless the true correlation between the two is zero, both X1 and X2 contain information about the true score for X1, here indicated as X1t, and both can be used in a regression equation to obtain predicted true X1t scores.

Information about X1t from the actual X1 scores is reflected in the reliability coefficient. From X2 scores it is found in the deviations of X2 values from values predicted by the regression of X2 on X1 within the experimental and control groups separately.

They present the following general formula (p. 71), expressed here in terms of X1 and X2:

$$\hat{X}1t = r_{XX}X1 \quad + \quad B_{X1t(X1 \cdot X2)}(X1-X2) \quad + \quad (1 - r_{XX})\,\overline{X1}$$

Where X1·X2 is their notation to represent residual scores formed from predicting X2 from X1 and the final expression is an adjustment of the mean of the group in terms of the lack of correlation between true and obtained scores, i.e. unreliability. Apparently, the reliability coefficient used in the equation would also be calculated within each group although the authors do not specifically say so.

One can pool the groups to obtain single "within-group" values for the parameters in the above equation but the estimates will be better when calculated separately within groups. This is especially true for groups not formed randomly because the true score distributions within each tend not to be the same; this implies that the same observed score, say X1, has a different true score, X1t, depending upon the group.

Cronbach and Furby go on to say that other relevant T1 variables should also be included in the true score estimate, the major limitation being the sample size used in the regressions. Wiley and Harnischfeger (1973) qualify that recommendation advocating that such other variables should be used only if one can defend, through a causal model, that they are theoretically direct determinants of the X2 variable. Their impression is that Cronbach and Furby would throw a whole pile of T1 background and other variables into the analysis indiscriminately in the hopes of reducing error.

Cronbach and Furby provide a method of calculating true score variances and then inputing these into regressions rather than using raw scores. They also distinguish between linked and unlinked T1-T2 measures and furnish adjustments that need to be made respectively.

Although my major concern was initially with the Cronbach-Furby approach, I want to discuss two others with the same underlying focus.

Wiley and Hornik Method

Wiley and Hornik (1973), to get accurate true score estimates, developed a measurement model to deal with errors in panel data but it relies on having more than one measure of each variable at each point in time. The two measures for any single variable will reflect the same true score variance but different error variances. Calling classical test theory into play, they make certain assumptions about the independence and additivity of variance components. Their use of cross-time and

alternate observed measures allows for the estimation of unobserved true and error variance in the variables. Once the true variances and covariances of the variables have been calculated they can be used to compute regression weights for the relationships among variables over time.

Hannan and Others

Wiley and Hornik refer to more "optimal" methods for using multi-wave longitudinal data. They are more optimal in the sense that they reduce the standard error around the estimates because the calculations involved employ more of the pieces of variance information that are available.

The references they cite, particularly for the relevant computer program, are the same as those John Meyer gave me as he talked about confirmatory factor analysis. Meyer's suggestion was to use it only if we found statistical significance with a technique developed by Michael Hannan and Alice Young (n.d.). Their method was an attempt to pool variance information about the variables of interest across all waves in an attempt to get a more accurate handle on statistical significance; like others, their intent was to reduce the amount of error variance to get an accurate estimate of true score relationships. The Hannan-Young method and confirmatory factor analysis require complicated estimation procedures which cannot be performed with least squares regression analyses.

Problems with the Strategies

MITT's concern with the Hannan-Young and Confirmatory Factor Analysis techniques was partly in their questionable appropriateness to the MITT design. Hannan and Young advocate a particular model whose use is constrained by several assumptions in the user's data (most of which we found difficult to grasp) and which itself has not been well tested. Their report is only one which does appear to lend support to the model's utility, but under what circumstances I'm not sure.

The Wiley-Hornik method is simpler to use but does not seem to fit our design either. Furthermore, for a study of MITT's magnitude the required calculations appear quite laborious.

Of any of the methods, the Cronbach-Furby and the Correction for Attenuation seem the most straightforward at first glance. Yet, certain characteristics of the MITT study leave these open to question also.

Before discussing them, it may first be worth noting that the variability of the sample under study affects the reliability -- one reason for requiring large random samples in reliability studies. One way of expressing the coefficient is as follows:

$$r_{XX} = 1 - \frac{\sigma^2_{meas}}{\sigma^2_X}$$

The variance of the errors of measurement (σ^2_{meas}) is considered to be approximately independent of the variance of the obtained scores (σ_X^2) and is therefore conceptually regarded as a fixed characteristic of the instrument regardless of the sample being studied. As the variance of the sample increases then, the ratio $\frac{\sigma^2_{meas}}{\sigma_X^2}$ will decrease and the reliability r_{XXX} will increase; as the variance decreases the ratio will increase and the reliability decrease.

Nunnally (1967) notes that a low reliability for an instrument will make detection of statistical significance difficult; when the standard error of measurement and the standard deviation of the variable in the sample are approximately equal he claims it is hopeless to investigate the variable.

My concerns center around four major points;

(1) I question the suitability of the formulas for MITT data; at least, I am not sure the models were developed with our type of instrumentation and unit of analysis problems in mind. The Cronbach-Furby and the Attenuation Correction methods have their focus on research using tests of abilities, I.Q., and personality traits which themselves have a tradition of being extensively researched for reliability and validity on large random samples of subjects. Consequently, the reliabilities of such measures are generally acknowledged as being stable and accurate and not affected by the variances of research samples in which they are used.

MITT deals with different phenomena; many of the types of variables involve opinions about others, perceptions about the school, some perceptions of self, and descriptions of behavior. None of the instruments received the extensive attention that trait and ability measures have traditionally received nor were their reliabilities calculated on exceedingly large random samples. This is not to say they are no good; it just questions the amount of confidence we can place in the accuracy of the reliabilities that would be used to estimate true scores. Questionable reliabilities would guarantee us nothing much better than questionable true score estimates.

(2) The accuracy of the reliabilities relates to a more difficult problem: How is a reliability to be computed?

a) If it is to be computed for each group (experimental vs. control) then it will change in accord with differences in the variance of each group even though a reliability, like the standard error of measurement, is conceptually a characteristic of an instrument independent of any sample of subjects.

b) Since we are involved in a school level analysis can we justifiably use reliabilities based on individuals when our unit of analysis is the school? This is part of an aggregation problem discussed by Hannan (1971) which maintains that aggregated scores measure a theoretically different variable than the unaggregated scores; perhaps, reliabilities should be based on school scores rather than individual scores.

(3) The Cronbach-Furby method includes a term consisting of residuals, deviations of X2 from predicted X1 scores as calculated

within each group. The logic of using posttest to predict pretest is that,

> Within a treatment, persons higher on the posttest than others having the same observed pretest score tend to be those for whom the true pretest score is higher than the observed score. (p. 72.)

This may be applicable for ability and trait measures but I'm not sure how accurately it models a lot of relationships with the MITT measures. For example, we may find that an experimental school has more classroom communication than another experimental school but that both have the same amount of communication at the previous wave. The contention of Cronbach and Furby is that the first school should theoretically have a higher communication level at the previous wave also.

(4) Finally, whether one employs variances or raw score regressions to compute true scores according to the Cronbach-Furby approach, he still faces the prospect of error in the variances due to small sample size. This is especially the case if true scores for variables must be estimated within each group.

In summary, we found no guarantee that any of the methods above would provide us accurate estimates of true scores. The arguments for use of the methods make sense but we were not sure how accurately the models for handling error reflect the characteristics of our design and variables. We do think a study to employ the models on the MITT data is something that could be written as a proposal itself. My recommendation at this point is to proceed as we have been.

APPENDIX C:

Rudiments of Path Analysis

Although path analysis cannot prove causality it can lend more or less confidence to a postulated model that describes the relationships one expects to find among his variables. The crucial, and probably most beneficial, aspect of the method is that it requires a clear specification of the causal model; the more ambiguous the variables and their relationships, the less confidence one can place in the analysis (See Appendix C).

The basic inferential tool crucial to path analysis is multiple linear regression which allows one to examine the magnitudes and directions of "direct" effects and their statistical significance while controlling for mutual influences among independent variables. The hypothesized causal model itself can be represented by a set of multiple linear regression equations.

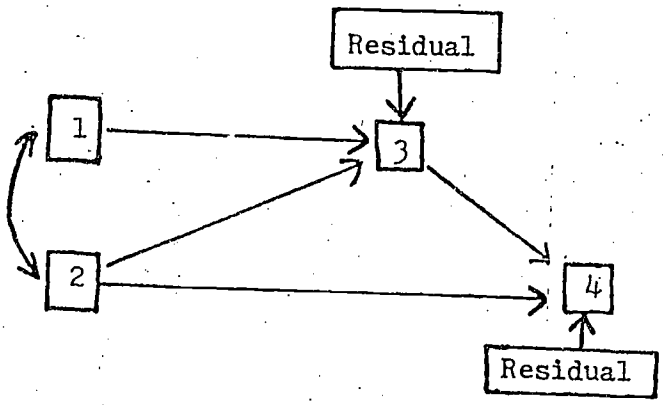
Because variables in behavioral science research are often expressed in arbitrary scales, not much substantive information about a path analytic model is conveyed by non-standardized regression weights, which specify that a 1.0-point change in the independent variable causes b points change in the dependent variable. This is because the different scale ranges of the independent variables obscure the importance of different variables relative to one another when the nonstandardized b-weights are used.

For this reason the standardized regression weights, B's or betas, are used as path coefficients to represent the direct effect of independent on dependent variables. Each coefficient estimates the amount of change in standard deviation units of the dependent variable that is produced by a 1-standard deviation change in the respective independent variable (Amick and Walberg, 1975).

The use of multiple linear regression in path analysis focuses upon explanation rather than classical prediction. The investigator desires not only to explain a substantial proportion of variance in the dependent variables but also to assess the relative importance of theoretically relevant independent variables.

The first step in formally analyzing data by path analysis is to explicitly specify a presumed unidirectional (recursive) causal ordering among the set of variables of interest. This model purports that the correlation between any two variables, except for those not causally determined by any other variables in the model, can be decomposed into a term representing the direct effect of one on the other plus a series of other terms representing the indirect effects. The indirect effects reflect portions of the correlation explained by spurious and/or mediated relationships.

The following diagram represents an example of a causal model with variables ordered from 1 to 4.



Variables #1 and #2 have no hypothesized causal determinants among the selected variables and therefore, the numbers signify only that they are different and not that one causally precedes the other. The two-headed arrow between them indicates that we cannot analyze their correlation. All other variables do have some hypothesized causes.

The diagram depicts what is called a recursive model because the causal flow is in one direction. Single-headed arrows between variables represent direct effects. Notice that some variables may each act as an independent variable and also as a dependent variable with respect to a subset of other variables in the model. Multiple-step paths showing variables acting through other variables to influence a dependent variable represent indirect effects. For example, the correlation between variables #2 and #4 is accounted for by a direct effect of #2 on #4 and a mediated effect of #2 acting through #3 which in turn affects #4. Because it is impossible to explain the total variation in any dependent variable completely by the designated independent variables, the residual variable is needed as a catch-all to account for all variance unexplained by the variables under scrutiny (Kerlinger and Pedhazur, 1973; Namboodiri et. al., 1975).

APPENDIX D:

Autocorrelations and Changes in Means

Usually, a longitudinal analysis will examine means at each wave to determine whether or not change occurs. In conjunction with them, the the autocorrelations can provide some useful information about the variation in variables and prevent hasty generalizations of the through-time trend in group means to each school comprising the group. If means increase or decrease one often tends to infer that the level of the variable in each school does so also; if the mean level of the variable remains the same, one may similarly infer that nothing has changed in the schools. But, the change in means is an index of group tendency not individual school variation. The group mean may show no change from one time to the next even though individual schools change drastically. The group mean may show a drastic jump or decrease even though some schools either do not change or change in the opposite direction.

The autocorrelation of a variable between two time points, however, can provide information that would help confirm or caution such a generalization. If the mean level changed but the autocorrelation were low, one would hesitate to generalize the trend to the majority of schools in the sample. If the mean levels changed and the autocorrelation were high, one would have confidence in generalizing. If the mean level remained the same and the autocorrelation were high, one would confidently generalize that the schools tended not to change.