THE USE OF WITHIN-GROUP SLOPES AS INDICES OF GROUP OUTCOMES

Leigh Burstein
M. David Miller
Robert L. Linn

CSE Report No. 171
1981

Center for the Study of Evaluation
Graduate School of Education, UCLA
Los Angeles, California 90024

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Abstract

The possibilities and problems in the use of within-group slopes of outcomes on inputs as indicators of substantive group effects are considered. Slopes are proposed as outcome measures which may reflect within-group processes in between-group analyses of multilevel data. Research on aptitude x treatment interactions, contextual effects and school effects provide a theoretical rationale for the proposed methodology. Data from the IEA Six Subject Survey are used to illustrate how a group-level analysis with slopes as outcomes might look. Finally, the statistical, empirical, substantive, and communication problems that arise from the use of slopes as outcomes are discussed.

Slopes As Outcomes

In recent years, there has been an increasing awareness that a thorough investigation of the effects of educational processes requires a multilevel examination of educational data (Burstein, 1980a, 1980c; Cronbach, 1976; Haney, 1974). Because of its multilevel (more precisely, hierarchical) organization, the effects of schooling on individual pupil performance can exist both between and within the levels of the educational system. Moreover, analyses at different levels address different questions and analyses conducted at a single level in such contexts have inherent problems.

Though choosing a unit of analysis dominated past discussions, expecially in program evaluation (cf. e.g., House, Glass, McLean, & Walker, 1978), current emphasis has shifted toward letting the choice of analytical model be dictated by the substantive processes under investigation (Burstein, 1980a, 1980b, 1980c; Burstein & Miller, 1978; Cronbach, 1976; Haney, 1974). That is, investigators are devoting greater attention to the development of adequate theories of educational processes and the determination of analytical methods for identifying the effects of such processes. These two activities are the basic elements in the specification of appropriate analytical models.

Within the domain of research on educational effects, the substantive processes in operation are functions of the characteristics of pupils (e.g., aptitude, previous exposure, motivation), characteristics of the classroom (e.g., instructional content and organization, peer

abilities and support, teacher style) and characteristics of the school (e.g., physical resources, academic atmosphere). Moreover, these substantive processes have collective (for the class or school as a whole) as well as individual effects (e.g., Burstein, 1976, 1980a, 1980b, 1980c; Burstein & Miller, 1978, 1980; Cronbach, 1976; Wiley, 1970). Given these features of multilevel educational data, the primary difficulties in proper model specification are the determination of the key substantive questions and the identification of evidence from the multiple levels that can potentially resolve them.

There are various aspects of the problem of proper model specification with multilevel data. On the one hand, there is a need for clearer conceptualization of the connections between properties of groups (ability level, cohesiveness) and the processes within groups (learning, interaction, participation). On the other hand, the special features of multilevel educational data call for special analytical methods designed for their examination.

This paper represents one attempt to mold analytical methodology to the special needs of multilevel data. Specifically, its purpose is to consider the possibilities and problems in the use of within-group slopes of outcomes on inputs as alternative indicators of substantive educational effects.

Theoretical Rationale

For the remainder of the paper, we restrict our attention to various types of field studies of educational effects and assume that the substantive questions of interest warrant group-level (classroom or school) analyses. For example, an investigator might be interested in performance

differences of classrooms which vary in their degree of structuring, emphasis on basic skills, or emphasis on cooperation. In such cases, it is possible to view the sampled classrooms as alternative "treatments" which vary along multiple dimensions and examine the relationships between a class's scores on the various dimensions and its outcomes. Much of the process-product research on teacher effectiveness (e.g., Anderson, Evertson, & Brophy, 1978; Brophy & Evertson, 1974), work on education production functions (e.g., Averch, Carroll, Donaldson, Kiesling, & Pincus, 1972), and school effects research (e.g., Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld, & York, 1966; Comber & Keeves, 1973) fits the above description. To some degree, large-scale evaluations of educational interventions such as Project Follow Through (House, et al., 1978; Stebbins, St. Pierre, Proper, Andersen, & Cerva, 1977) can be viewed in a similar fashion (Rogosa, 1978).

Regardless of the type of field study being conducted, once it has been determined that the substantive questions of interest warrant examination of differences among groups, the type of between-group effects one expects to find remain to be specified. While analyses of the relationships between "treatment" dimensions and the mean outcomes of groups often provide useful information, important differences in within-group processes may be obscured. These within-group processes may arise due to group composition (e.g., ability level and mixture affecting participation patterns), differential allocation of instructional resources among the members of the group (e.g., the grouping and pacing features of reading instruction), or differential reactions of group members to the same instructional treatment (aptitude-treatment interactions).

If important group-to-group differences in within-group processes exist, then the use of group means as the only indicator of group outcomes

can result in misleading or, at least, incomplete estimates of group (teacher/class/school/treatment) effects. In such cases, other indices of group outcomes such as the standard deviation (Brown & Saks, 1975; Klitgaard, 1975; Lohnes, 1972) should be considered.

Our interest in alternative measures of group outcomes has concentrated on the properties of the within-group slopes from the regression of outcomes on input (Burstein, 1976, 1980a, 1980b; Burstein, Linn, & Capell, 1978; Burstein & Miller, 1978, 1980). Within-group slopes may be viewed as group-level indicators of within-group processes. Moreover, differences in slopes across groups can be the result of substantive educational effects.

That is, we suggest that variation in slopes across groups can reflect the influence of group characteristics such as the level and distribution of instructional resources. For example, the relationship of ability to achievement within classes with educational "treatments" involving high levels of emphasis on grouping and pacing may differ markedly from classes with low levels on these "treatment" characteristics. Under circumstances where classrooms differ on what are perceived to be important instructional characteristics, it seems logical to inquire about whether, ceteris paribus, these differences are systematically related to variation in the within-class relationship of ability to achievement. If such relationships exist, then it can be argued that the within-group slope, a group-level outcome, varies as a function of a within-group process. Later on we provide some caveats about attempts to account for slope differences.

To our knowledge the specific features of our approach for analyzing variation in within-groups slopes have not been previously investigated

in educational research. Interest in the potential substantive importance of heterogeneity of within-group regressions is, however, not new. Slope heterogeneity is studied in psychological research on aptitude-treatment interactions, in sociological research on context effects, and in work on interactive effects of opportunity to learn. Before describing our own conceptual and empirical work on slopes as indicators of group outcomes, we review briefly the literature on these topics.

Heterogeneous Slopes as Aptitude-Treatment Interactions

Research on aptitude-treatment interactions (ATI; Berliner & Cahen, 1974; Cronbach, 1976; Cronbach & Snow, 1977; Cronbach & Webb, 1975; Snow, 1976) provide the original impetus for our examination of within-group slopes. The logic of ATI research is built on the substantive significance of differences in within-treatment regressions. For example, it may be theorized that a highly structured presentation might lead to a weaker relationship between entering aptitude and final achievement than would a treatment with less structure; or a competitive treatment would lead to a stronger relationship than a cooperative treatment.[1]

ATI logic can be carried to the level of the individual groups (classrooms, schools). Each classroom becomes a treatment whose characteristics may be measured along several dimensions. If classrooms contain pupils with similar distributions of entering characteristics (e.g., comparable pretest and aptitude distributions), then differences in within-class slopes would be anticipated on the basis of knowledge of differences in instructional methods and resources. For example, it might be hypothesized that there would be flatter slopes for classrooms in which the teachers target instruction to improve the performance of lower-ability students than in classrooms where students are allowed to learn at their own rate.

There are several examples of the consideration of heterogeneous within-group regressions (nested within treatments) in the recent ATI literature (Corno, 1979; Cronbach, 1976; Cronbach & Snow, 1977; Cronbach & Webb, 1975; Greene, 1976, 1980; Gustafsson, 1978; Snow, 1976). For example, in their multilevel reanalysis of the Anderson (1941) data on drill vs. meaningful instruction in arithmetic, Cronbach and Webb (1975) found that class-by-class regressions (N=18 classes) varied greatly. However, they considered the overall proportion of variation due to within-class regressions to be small (4.1 percent for the drill treatment and 6.9 percent for the meaningful treatment). Moreover, several unusual slopes could be traced to the effects of outliers (anomalous students within classes). Cronbach (1976) reached similar conclusions in a reanalysis of selected data from the Cooperative Reading Study (Bond & Dykstra, 1967).

Greene (1976, 1980) investigated the effects of choice (when, how long, in what sequence) and no-choice treatments on learning from workbook lessons. Both treatments were randomly assigned to half of the students in nine fourth and fifth grade classes. The heterogeneity of within half-class regressions of outcomes on general ability is striking (see Figure 2, p. 84 in Snow (1976) and Figure 1, p. 298 in Greene (1980)). While acknowledging the limited stability of slopes based on approximately 12 observations, both Greene and Snow point out notable within-class differences which are consistent with theories about the appropriate aptitude-treatment match.

In the studies cited above, the examination of heterogeneous within-group regressions was only of secondary interest. Treatments were considered to be discrete; a class is in either treatment A or treatment B.

Variability in slopes across classrooms within treatments represented either a nuisance or food for thought.

Despite its theoretical soundness, the practical mechanics of extending current lines of ATI inquiry to the consideration of classrooms as distinct treatments which vary quasi-continuously along a number of dimensions are complicated. Each new treatment dimension and for that matter, aptitude dimension, forces the investigator into the consideration of a higher-order interaction (Cronbach, 1975). Though extention via the general linear model is seemingly straightforward, current methods of conceptualizing and analyzing higher-order ATI's lack substantive and statistical power. The requirements for valid and reliable indicators of treatment dimensions may be too difficult to surmount given the present state of knowledge in this area.

Slope Heterogeneity and Context Effects

A concern for the heterogeneity of within-group relationships is fundamental to certain approaches to contextual analysis in sociology and political science (Boyd & Iverson, 1979; Valkonen, 1969). Contextual analysis is the study of the effects of properties of groups or collectives on individuals (Lazarsfeld & Menzel, 1961).

In its extended form (cf. Boyd & Iverson, 1979), the basic contextual model specifies that an individual-level dependent variable (Y_{ij}) is a function of individual-level explanatory variables (X_{ij}), their group-level counterparts ($\bar{X}_{i.}$) and the interaction between individual-level and group-level variables ($X_{ij}\bar{X}_{i.}$):

$$Y_{ij} = a + \beta_1 X_{ij} + \beta_2 \bar{X}_{i.} + \beta_3 (X_{ij}\bar{X}_{i.}) + \varepsilon_{ij} \qquad (1)$$

A typical contextual analysis interpretation is that a nonzero value of β_3 (or a significant heterogeneity of regression in an analysis of

covariance) implies that the relationship between X and Y varies as a function of the level of the group on the explanatory variable.

More generally, Boyd and Iverson (1979) suggest that the connections between within-group relationships and specific properties of groups be investigated in the two regression equations:

$$\alpha_i = F(\underset{\sim}{W}_i) + u_i \qquad (2)$$

$$\beta_i = F(\underset{\sim}{W}_i) + v_i \ . \qquad (3)$$

In (2) and (3), α_i and β_i are the intercept and slope from the within-group regressions of Y_{ij} on X_{ij} and the $\underset{\sim}{W}_i$ measure certain properties of groups and $F(\underset{\sim}{W}_i)$ denotes an unspecified functional form of the $\underset{\sim}{W}_i$.

Boyd and Iverson (1979) consider in detail the case where $\underset{\sim}{W}_i = \bar{X}_{i.}$ and illustrate how various combinations of individual, group, and interaction effects give rise to specific effect estimates in group-level analyses of mean outcomes. They also describe how their form of contextual analysis would proceed when group variables other than those based on group means are used (Section 3.4).

It is clear that the Boyd-Iverson approach to contextual analysis recognizes the integral role of within-group slopes in the examination of group properties and processes. However, their treatment is purely didactic. At present there are no actual empirical examples of how such an analysis might look. Moreover, the implication of the specific form of heterogeneity reflected in (1) needs to be addressed.

Interactive Schooling Effects of Opportunity to Learn

Though approached from a different methodological perspective, Sorenson and Hallinan's (1977) reconceptualization of school effects also embodies underlying heterogeneity of relationships across schools. They

view learning as a time-dependent process wherein the variation in the amount of learning achieved is a function of three concepts -- ability, effort, and opportunity to learn. Sorenson and Hallinan offer a specification for the interrelations among these three concepts in which the effects of ability and effort on learning are constrained by the opportunity to learn. They carry the reasoning a step further to suggest that between-school variation in opportunity to learn can lead to heterogeneity among schools in the relationship of ability and effort to learning.

Sorenson and Hallinan's proposed specification is a differential equation model for change in achievement. However, they point out that a reasonable representation of their conception of the learning process can be found through the estimation, separately for each school (classroom) (Sorenson & Hallinan, 1977, p. 278), of the regression of achievement after exposure to a learning process of length t on initial achievement and individual characteristics representing ability and effort.

According to Sorenson and Hallinan (p. 278), variation among schools in the relationship of achievement at time t to initial achievement provides information on the variation in opportunity to learn. Thus, they anticipate differences in within-group slopes (of post-achievement on pre-achievement) which would reflect the interactive effects of schooling which arise through differences in opportunities for learning. To emphasize further their perspective, Sorenson and Hallinan focus directly on slope heterogeneity in their empirical example.

Examining Slopes as Outcomes

With the exception of Sorenson and Hallinan (1977), the research described above has involved natural extensions of the general linear model to incorporate hypothesized heterogeneous within-group regressions. The multilevel ATI work to date relies mainly on descriptions and discussion of plots of within-group regressions accompanied by information on variance decomposition (e.g., specific within-class variation vs. pooled within-class variation). And, while the modeling of the within-group regressions in (2) and (3) is an integral part of the Boyd-Iverson contextual analysis, this activity is viewed as secondary to the examination of the general model (equation (1)) and its associated variance decomposition.

Our emphasis departs from the work cited above in that the within-group slope becomes an additional integral variable whose variation is to be explained. That is, we examine the use of within-group slopes of outcomes on inputs as a criterion measure in studies of educational effects. Wiley (1970) was apparently the first to suggest this strategy. As part of his argument that the collectivity (class, school, etc.) is the appropriate unit of analysis in educational evaluation, he commented that the focus on the mean level of achievement of the collectivity may be too narrow. Wiley suggested that the moments of the achievement distribution, contrasts between sub-populations and regression coefficients might be used as criterion measures for evaluating the differential effect of instructional treatments on individual pupils.

Our reason for considering slopes as outcomes is that there may be instructional effects on the within-group regression of outcomes on

input, whether there are instructional effects on group mean performance or not. If slope effects are present, the analysis should attempt to isolate instructional process and practice variables that are associated with slope variation. If such variables can be found and alternative explanations cannot be ruled out, then variation in slopes becomes an important source of information for researchers and policy makers, expecially when considered along with effects on other group-level outcomes.

In practice, our empirical investigations have treated within-group slopes as one of several outcomes in a between-group analysis. In the example to follow, we looked at three group-level models:

$$\bar{Y}_i = \alpha_1 + \underset{\sim}{\gamma}_{\bar{X}}\underset{\sim}{\bar{X}}_{i.} + \underset{\sim}{\gamma}_{S}\underset{\sim}{S}_{i.} + \varepsilon_1 \qquad (4a)$$

$$\hat{\beta}_i = \alpha_2 + \underset{\sim}{\theta}_{\bar{X}}\underset{\sim}{\bar{X}}_{i.} + \underset{\sim}{\theta}_{S}\underset{\sim}{S}_{i.} + \varepsilon_2 \qquad (4b)$$

$$\hat{\sigma}_{u_i} = \alpha_3 + \underset{\sim}{\delta}_{\bar{X}}\underset{\sim}{\bar{X}}_{i.} + \underset{\sim}{\delta}_{S}\underset{\sim}{S}_{i.} + \varepsilon_3 \qquad (4c)$$

In the above $\underset{\sim}{\bar{X}}_i$ and $\underset{\sim}{S}_i$ denote vectors of input (background) and schooling characteristics, respectively. $\bar{Y}_i$ is the mean of the distribution of outcome scores within group i. $\hat{\beta}_i$ is the slope from the regression of outcome on input (in this case, a measure of verbal ability) in group i. $\hat{\sigma}_{u_i}$ is the standard error of estimate from the regression of outcome on input within group i.

The γ's, θ's, and δ's are coefficients from the three regression equations based on group-level data. The $\underset{\sim}{\theta}$'s presumably reflect any systematic effects of schooling characteristics on slopes. This interpretation of $\underset{\sim}{\theta}$'s is directly relevant to an elaboration of the components of educational effects on pupil outcomes.

The actual analysis for slopes is a two-step procedure whereby the within-group slope is estimated separately for each group before

being used as an outcome measure in equation (4b). There are multiple background factors in the empirical analysis. However, we concentrate strictly on the regression of outcome on a single input (verbal ability) for the sake of explanation and because we believe that the slope differences for other background variables are inconsequential after verbal ability is controlled.

The choice of standard errors of estimate as a group outcome represents a departure from earlier use of standard deviations as group outcomes (e.g., Brown & Saks, 1975; Burstein, 1980a, 1980b). However, our focus on heterogeneous within-group slopes dictates against considering the standard deviation as an indicator. Groups with similar distributions (e.g., standard deviations) of entering student characteristics would be expected to have different outcome distributions if instructional practices led to variation in slopes. That is, heterogeneity of regression across classes would result in the heterogeneity in standard deviations of outcomes across groups with similar standard deviations on entering characteristics. Thus, slopes and standard deviations are likely to be correlated (as they are here, see Table 1). The standard error of estimates, however, can serve as a measure of outcome variation across groups that is not likely to be related to the slopes. As a consequence, the variables that predict variation across groups in the standard error of estimate should be either background conditions not adequately reflected in the slopes or schooling characteristics that influence performance in a manner not systematically related to entering characteristics.

In earlier work (Burstein et al., 1978), hypothetical data were generated to examine the effects of heterogeneous within-group slopes on

the results from several analytical models for identifying educational effects. Under the conditions studied, when differences in within-group slopes were systematically related to teacher/class characteristics, the slopes-as-outcome analysis (equation 4b) exhibited desirable properties. This analysis, when conducted in conjunction with class-mean analysis (equation 4a), identified the direction and, to some degree, the severity of bias in estimating teacher effects on class mean outcomes. It also suggested that such an analysis might help to disentangle the multiple effects of schooling. Below, we carry this activity a step further by examining the effects of schooling on all three group outcomes (means, slopes, standard errors of estimate) in a specific empirical example.

An Empirical Example

To make the above discussion more concrete, we elaborate on an empirical example using IEA science data on U.S. fourteen year-olds (see Comber & Keeves, 1973 for a description of the original study) which has been presented previously in somewhat different forms (Burstein, 1980a, 1980b; Burstein & Miller, 1980). In this example, a school-level analysis (N=107 schools) of factors affecting science test performance is considered. The explanatory variables include ascribed background characteristics (sex, socioeconomic variables), a concurrent measure of verbal ability and two characteristics of science instruction (instructional approach and present exposure to science) (see Table 1 for a description of variables). All explanatory variables are school-level averages of individual student responses. Thus, it is possible that the group-level effects discussed here are in part simple aggregations of individual effects (Alwin, 1977; Boyd & Iverson, 1979; Firebaugh, 1978).

The science test was used as the outcome measure and the verbal ability test was treated as a proxy for the input measure. That is, science scores were regressed on verbal ability scores within each school. Longitudinal data with pre and post instruction measures would clearly be preferable to the cross-sectional data that were used in this example. The verbal ability score is only a crude proxy for a pre-instruction measure of student ability. It is adequate, however, for the illustrative purposes of this paper.

The mean science score, the slope of the regression of science on verbal ability, and the standard error of estimate from the regression were then taken as the three descriptors of science achievement outcomes for a school. These three outcome indices were then regressed on the background and explanatory variables. Table 1 provides descriptive data on the variables included in the regression analysis. Note that within-school slopes are strongly related to the standard deviation of science scores (r=.52), but are weakly related to school mean science score (.18) and the standard error of estimate (.10).

Table 1

Table 2 presents the school-level regressions of means, slopes, and standard errors of estimates on school means on background and schooling characteristics. The same set of explanatory variables has been used in all three analyses for comparison purposes though in theory, different characteristics could be expected to influence the different indicators of group outcomes.

Table 2

Substantive interpretations of data such as those in Table 2 would presumably deal with each outcome in turn. In the present example, the model for explaining variation in school mean science performance displays a relatively good fit. The coefficients for the two schooling characteristics suggest that for schools at a given level on background characteristics, performance is higher when students are receiving more science instruction and when science instruction emphasizes discovery methods.

The examination of the effects of schooling characteristics on within-school slopes (table discussed below) and standard errors of estimate does appear to elaborate how a school's science performance is affected. The coefficients from the standard error of estimate model suggest that variation of individual performance about the within-school regression is greater in schools in which students report a large number of books in the home and high exposure to science instruction.

We can think of at least two mechanisms that might account for the science instruction effects on the standard errors of estimate. First, it is possible that in schools with more opportunity for exposure to science instruction, students with similar verbal skills may vary in the degree to which they forego or take full advantage of increased opportunities. As a consequence, there would be large differences in science performance of students with similar verbal abilities simply due to differences in actual versus possible exposure to science instruction. Alternatively, schools with high levels of science instruction may require students with similar verbal skills to receive more science instruction than in schools with lower levels of science instruction regardless of students' interest in science or aptitude for science study. In such schools, variation in

science scores for students with similar verbal abilities would be expected to the degree that science aptitude and interest mediate performance. Under either condition (differential opportunity, or low aptitude and/or interest with similar opportunity), it is reasonable to expect substantial variation of science performance among students with similar verbal abilities in schools offering greater opportunities.

The effects of the schooling variables instructional approach and amount of science instruction on the slopes provide an indication of which students benefit practices. The relationship of a student's verbal score to their science score is apparently stronger in schools with high (verbal) ability students, with a high proportion of male students, with more exposure to science and with greater emphasis on discovery approaches. To grossly simplify matters, given two schools with the same sex ratio and overall mean verbal ability score, the difference in performance between a student with a lower verbal ability and one with higher verbal ability would be expected to be greater in the school offering more science instruction and utilizing a discovery approach.

To highlight the contrasts, expected differences in science scores were estimated for hypothetical students at various levels of verbal ability in schools with below average, average, and above average levels of discovery approach to instruction (EXPLORE) and amount of science instruction (SCIINST) and at the average on all other variables (see Table 3). For the extreme cases in the table [(+1,+1) and (-1,-1)], the difference in science scores between a school's lower and higher verbal ability students is expected to be 1/2 of a standard deviation ((9.7 - 5.7)/8.1).

Table 3

The results in Tables 2 and 3 suggest that greater opportunities for science study and a discovery emphasis in science instruction do lead to higher pupil performance on the average. Instruction which emphasizes student self-direction (selection) of learning goals and inductive problem-solving tends to magnify pre-existing differences in pupil skills. Higher ability students tend to make more appropriate choices and learn better under these conditions than lower ability students. The steeper within-group slopes with greater opportunities for exposure to science instruction and with greater emphasis on individual exploration are consistent with results from research on informal/open individually guided/less structured instruction (e.g., Bennett, 1976; Peterson, 1977; Stebbins et al., 1977).

The above discussion probably overemphasizes the practical impact of the schooling characteristics. In order to gain a better understanding of the source of the impact, the within-school regressions for the ten schools with the highest EXPLORE and SCIINST scores and the ten schools with the lowest are examined. Figures 1 and 2 contain lines showing the regression of science scores on verbal ability for the high and low EXPLORE, and high and low SCIINST schools, respectively. The endpoints of the line for each school coincide with points plus or minus one within-school standard deviation from the school's mean on verbal ability.

Figures 1 and 2

There appear to be some discernible patterns. In Figure 1, four out of the five schools with the highest slopes (E, K, M, L, and P) were high EXPLORE schools while all five of the lowest slopes (A, B, D, G, H) were low EXPLORE schools. The high EXPLORE schools had a mean slope of .93 with a standard deviation of .40 while the low EXPLORE schools had a mean slope of .48 with a standard deviation of .44, a statistically significant difference ($p < .03$). The differences are most marked for schools with average mean verbal ability.

The plots for schools with high and low levels of science instruction are less clear than those for EXPLORE schools (the EXPLORE and SCIINST schools are not the same though they overlap). There are several schools with low verbal ability and high science exposure (e.g., K, P, T) or high verbal and low science exposure (e.g., H, I), so contrasting exposure at a given level of ability is less informative. The within-school slopes in low science instruction (SCIINST) schools tend to increase with ability while the slopes in the high instruction schools seem to vary less systematically. The mean and standard deviation of the slope distributions were .87 and .34, respectively, for high SCIINST schools and .65 and .43 for low SCIINST schools, a statistically nonsignificant difference ($p < .22$). Clearly, there is a need for a more fine-grained look at schools with specific combinations of EXPLORE, SCIINST, and verbal ability.

Problems with Slopes as Outcomes

Despite their theoretical and empirical appeal, the use of within-group slopes as outcomes is fraught with problems. The problems cover

a broad research universe: statistical, empirical, substantive, and communicative. We briefly discuss each type below.

Statistical Problems

The mathematical properties of slopes as outcomes are not well understood. We are essentially treating the within-group slopes as a random variable with an unknown underlying distribution. Our logic is somewhat related to econometric work on random coefficients regression models (e.g., Akkina, 1974; Maddala, 1977; Swamy, 1970) though econometricians typically deal with the case where slope variation is a random variable unrelated to the explanatory variables in the model (for an exception, see Hanushek, 1974).

The criticism that within-group slopes should not be treated as random variables is troubling, but certainly not fatal. There are too many instances in behavioral research where sensible analytical work has been conducted without mathematical confirmation of the appropriateness of the distributional assumptions in the measurement of a crucial variable. Any score which is a simple sum of other scores is also subject to uncertainties. The final line of defense against the statistical criticism is that like any other measure of unknown properties, it is necessary to have a sound theoretical rationale for using it, to demonstrate its empirical utility, and seek to identify and disconfirm any counter-interpretations to theoretical and empirical evidence.

Empirical Problems

The empirical problems with studies of slopes as outcomes are generally the same as with any investigation of regression models. Group-to-group variation in slopes are notoriously sensitive to inadequacies and anomalies in the data. In general, regression coefficients are

strongly affected by measurement errors in the regressors, ceiling (floor) effects, outliers, small numbers of observations, and multicollinearity. In fact, some researchers view unusual group slopes as possible indications of outliers or ceiling effects especially when generated from data on classrooms.

We have no quarrel with these empirical concerns about slopes. In fact, investigators who wish to treat slopes as outcomes should examine the scatterplots and descriptive statistics for the individual classes or schools for outliers and ceiling effects as an essential precautionary measure. Outliers and floor and ceiling effects were excluded as threats to our interpretation of the IEA data. In most cases, the slope accurately characterized the bivariate distribution of a school's science and verbal scores. And, while sample sizes were relatively small in some schools (as low as 10), there was no clear relationship between slopes and sample size or between standard errors of estimate and size. Any attenuation problems due to measurement errors in the regressors is minor since the psychometric properties of the verbal score used above are very good across the whole sample (internal consistency coefficients above .9). Moreover, there is no evidence that measurement error problems are more severe in some schools than in others which would have to be the case to challenge any identified effects on slopes as outcomes.

Substantive Problems

In earlier sections on the theoretical rationale for examining slope heterogeneity, we focussed on schooling characteristics (instructional approach, etc.) as the source of slope differences. Realistically, such interpretations are reasonable only for groups with comparable distributions of entering characteristics. It is highly likely that slope heterogeneity

in studies of naturally occurring educational groups can be more readily explained by selection effects (Alwin, 1976; Burstein, 1980a; Cronbach, 1976). Typically, classrooms and schools vary in the mechanisms which guided their formation (community wealth, pupil ability, etc.) and in their composition of student skills, background and attitudes. The flip-side of the ATI coin is that one can expect a different array of outcomes from a single treatment for classrooms (schools) which vary in their student composition. Heterogeneous vs. homogeneous ability grouping and high ability vs. low ability combinations would certainly lead one to expect different treatment outcomes and would itself be of substantive interest (Webb, 1980).

The analyst needs to be keenly aware of selection and composition at every stage of a multilevel investigation. In the present example, composition effects on slopes as measured by school means on verbal ability are certainly strong. The effects of the heterogeneity of verbal ability within the school on the slope are weaker, but significant nonetheless. However, composition effects as measured did not wipe out the more substantively interesting effects of science instruction characteristics.

Another substantive problem arises when the various indicators of group-level outcomes are highly correlated. Parsimony alone would argue that precedence should be given to simpler explanations. For example, one might argue that analyses of school means captures all of the meaningfully interpretable effects and presumed effects on slopes and standard errors of estimate are actually redundant with effects on means. Again, however, finding interpretable differences in patterns of effects across indicators is the best way to make a case

for separate examinations of other indicators besides group means. In our present example, we feel fairly confident that the effects on slopes cannot be solely explained by effects on group means. When a school's outcome mean and standard deviation are included as explanatory variables in the model with slopes as outcomes, the significant effects of EXPLORE and SCIINST are only marginally altered even though the overall proportion of explained variation is more than doubled.

Communication Problems

Communication problems refer to the whole class of difficulties in presenting a theory and describing research results in a manner that others will understand. This is a difficult task in multilevel analysis models, especially those which try to capture within-group phenomena by examining the antecedents of slope heterogeneity. Even in the simplest cases, the reader is asked to envision patterns in the distributions of lines across groups and try to relate these patterns to characteristics of the education in the groups. Anyone uncomfortable with either ATI reasoning or the conceptual distinctions possible with multilevel data is bound to balk when asked to understand models which combine the two lines of thought.

We have no simple answer to the communication problem in research and evaluations which involve multilevel educational data. While little work on multilevel problems was done in educational research between Wiley's conference presentation (actually presented in 1967, but not published until 1970) and Haney's (1974) paper on Project Follow Through, there has been a virtual flood of interest in recent years, traceable mainly to Cronbach's (1976) report. While the level of cognizance of

multilevel problems is fairly high at present, in education as well as other social science research, time and experience are the keys to either the demise of the concerns or the bridging of the communications gap.

multilevel problems is fairly high at present, in education as well as other social science research, time and experience are the keys to either the demise of the concerns or the bridging of the communications gap.

NOTE

1. In the structured/unstructured comparisons, the presence of structure presumably benefits the less able student by providing additional tools to tackle the tasks, thereby reducing the dependence of performance on prior ability (Peterson, 1977). In the competitive/cooperative example, the competitive environment offers no incentive to the more able student to help the less able thereby excerbating pre-existing ability differences which presumably reflect prior competitiveness and motivation (Hanelin, 1978; Slavin, 1977).

References

Akkina, K.R. Application of random coefficient regression models to the aggregation problem. Econometrica, 42(1974):369-375.

Alwin, K.F. Assessing school effects: Some identities. Sociology of Education, 49(1976):294-303.

Anderson, G.L. A comparison of the outcomes of instruction under two theories of learning. Unpublished doctoral dissertation, University of Minnesota, 1941.

Anderson, L.M., Evertson, C.M., and Brophy, J.E. The first grade reading group study: Technical report of experimental effects and process-outcome relationships (R&D Reports No. 4070). Austin, TX: Research and Development Center for Teacher Education, 1978.

Averch, H., Carroll, S.J., Donaldson, T., Kiesling, H.J., and Pincus, J. How effective is schooling? A critical review and synthesis of research findings (R-956-PCSF/RC). Santa Monica, CA: The Rand Corporation, 1972.

Bennett, S.N. Teaching styles and pupil progress. Cambridge, Mass.: Harvard University Press, 1976.

Berliner, D.C., & Cahen, L.S. Trait-treatment interaction and learning. In F.N. Kerlinger (Ed.), Review of Research in Education, (1973): 58-94.

Bond, G.L., & Dykstra, R. The Cooperative Research Program in first-grade reading instruction. Reading Research Quarterly, 2(1967): 5-142.

Boyd, L.H., and Iverson, G. Contextual analysis: Concepts and statistical techniques. Belmont, CA: Wadsworth Publishing, 1979.

Brophy, J.E., and Evertson, C.M. Process-product correlations in the Texas teacher effectiveness project: Final report (R&D Report No. 4004). Austin, TX: The REsearch and Development Center for Teacher Education, 1974.

Brown, W., and Saks, D.H. The production and distribution of cognitive skills within schools. Journal of Political Economy, 83(1975): 571-593.

Burstein, L. Assessing the differences of between-group and individual regression coefficients. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 1976.

Burstein, L. The role of levels of analysis in the specification of educational effects. In R. Dreeben and J.A. Thomas (Eds.), Analysis of educational productivity. Volume I: Issues in microanalysis. Cambridge, Mass.: Ballinger Press, 1980, 119-190. (A)

Burstein, L. Analyzing multilevel educational data: The choice of an analytic model rather than a unit of analysis. In E. Baker and E. Quellmalz (Eds.), Design, analysis, and policy in testing and evaluation. Beverly Hills, CA: Sage Publications, 1980, 81-94. (B)

Burstein, L. Analysis of multilevel data in educational research and evaluation. In D. Berliner (Ed.), Review of Research in Education, Vol. 8. Washington, D.C.: American Educational Research Association, 1980, 158-233. (C)

Burstein, L., Linn, R.L., & Capell, F. Analyzing multilevel data in the presence of heterogeneous within-class regressions. Journal of Educational Statistics, 3(1978):347-383.

Burstein, L., & Miller, M.D. Alternative analytical models for identifying educational effects: Where are we? Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, Canada, March 1978.

Burstein, L., & Miller, M.D. Regression-based analysis of multilevel educational data. In R. Boruch, D.S. Cordray, & P.M. Wortman (Eds.), Secondary analysis: Policies and practices for improving applied social research. San Francisco, CA: Jossey Bass, Inc., in press.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, S., Weinfeld, F.D., and York, R.L. Equality of educational opportunity (2 Vols.). Office of Education, U.S. Department of Health, Education and Welfare, Washington, D.C.: U.S. Government Printing Office, 1966.

Comber, L.C., and Keeves, J.P. Science education in nineteen countries, international studies in evaluation (Vol. 1). Stockholm: Almqvist & Wiksell; and New York: Wiley, 1973.

Corno, L. A hierarchical analysis of selected naturally occurring aptitude-treatment interactions in the third grade. American Educational Research Journal, 16(1979):391-410.

Cronbach, L.J. (with assistance of J.E. Deken & N. Webb). Research on classroomsand schools: Formulation of questions, design, and analysis. Occasional Paper, Stanford Evaluation Consortium, Stanford, CA, July 1976.

Cronbach, L.J., and Snow, R.E. Aptitude and instructional methods. New York: Irvington, 1977.

Cronbach, L.J., and Webb, N. Between-class and within-class effects in a reported aptitude x treatment interaction: Reanalysis of a study by G.L. Anderson. Journal of Educational Psychology, 67(1975): 717-724.

Firebaugh, G. A rule for inferring individual-level relationships from aggregated data. American Sociological Review, 43(1978):557-571-572.

Greene, J.C. Choice behavior and its consequences for learning: An ATI study. Unpublished doctoral dissertation, Stanford University, 1976.

Greene, J.C. Individual and teacher/class effects in aptitude treatment studies. American Educational Research Journal, 17(1980):291-302.

Gustafsson, J.E. A note on class effects in aptitude x treatment interactions. Journal of Educational Psychology, 70(1978):142-146.

Hanelin, S.J. Learning, behavior, and attitudes under individual and group contingencies. Unpublished doctoral dissertation, University of California, Los Angeles, 1978.

Haney, W. Units of analysis issues in the evaluation of Project Follow Through. Unpublished report, Cambridge, Mass.: Huron Institute, 1974.

Hanushek, E.A. Efficient estimators for regressing regression coefficients. The American Statistician, 28(1974):66-67.

House, E.R., Glass, G.V., McLean, L.D., and Walker, D.F. No simple answer: Critique of the Follow Through evaluation. Harvard Educational Review, 48(1978):128-160.

Klitgaard, R.E. Going beyond the mean in educational evaluation. Public Policy, 23(1975):59-79.

Lazarsfeld, P.F., Menzel, H. On the relation between individual and collective properties. In A. Etzioni (Ed.), Complex organizations: A sociological reader. New York: Holt, Rhinehart, & Winston, 1961.

Lohnes, P. Statistical descriptors of school classes. American Educational Research Journal, 9(1972):547-556.

Maddala, G.S. The use of variance components models in pooling cross-section and time-series of cross-sections. Econometrica, 39(1971): 359-382.

Peterson, P.L. Interactive effects of student anxiety, achievement orientation, and teacher behavior on student achievement and attitude. Journal of Educational Psychology, 69(1977):779-792.

Rogosa, D. Politics, process, and pyramids. Journal of Educational Statistics, 31(1978):79-86.

Slavin, R.E. Classroom reward structure: An analytical and practical review. Review of Educational Research, 47(1977):633-650.

Snow, R.E. Learning and individual differences. In L.S. Shulman (Ed.), Review of Research in Education, Vol. 14. Itasca, IL: F.E. Peacock Publishers, 1976.

Sorenson, A., and Hallinan, M.T. A reconceptualization of school effects. Sociology of Education, 50(1977):273-289.

Stebbins, L.B., St. Pierre, R.G., Proper, E.C., Andersen, R.B., and Cerva, T.R. Education and experimentation: A planned variation model (Vol. IV-A, An evaluation of Project Follow Through). Boston: Abt Associates, April 1977.

Swamy, P.A.V.B. Efficient inference in a random coefficient regression model. Econometrica, 38(1970):311-323.

Valkonen, T. Individual structural effects in ecological research. In M. Dogan and S. Rokkan (Eds.), Quantitative ecological analysis in the social sciences. Cambridge, Mass.: The MIT Press, 1969.

Webb, N.M. Group process: The key to learning in groups. In K. Roberts and L. Burstein (Eds.), Issues in aggregation, Number 6, New directions for methodology of social and behavioral science, San Francisco: Jossey Bass, Inc., 1980, 77-88.

Wiley, D.E. Design and analysis of evaluation studies. In M.E. Wittrock and D.E. Wiley (Eds.), The evaluation of instruction: Issues and problems. New York: Holt, Rinehart & Winston, 1970.

Table 1. Descriptive statistics for slopes and other school-level variables in the IEA science data for the U.S. (N=107 schools).

Variables[a]	Slope	Science Mean	Science S.D.	Standard Error of Estimate	Verbal Mean	Verbal S.D.	Sex	Father's Occupation	Books in the Home	Instructional Approach	Science Instruction
Slope											
Science Mean	.18										
Science S.D.	.52	.34									
Standard Error of Estimate	.10	.30	.85								
Verbal Mean	.24	.78	.21	.16							
Verbal S.D.	-.12	.19	.35	.22	-.15						
Sex	-.17	-.26	-.18	-.10	-.11	-.05					
Father's Occupation	.07	.64	.26	.23	.53	.14	-.09				
Books in the Home	.13	.73	.35	.31	.58	.17	-.08	.67			
Instructional Approach	.25	.29	.29	.15	.17	.24	.01	.25	.29		
Science Instruction	.23	.03	.15	.16	-.05	-.11	.09	-.17	-.03	-.09	
Mean	.76	57.28	6.92	5.87	27.54	4.45	1.52	6.14	4.55	9.84	1.59
Standard Deviation	.38	4.43	1.49	1.35	2.41	1.00	.16	1.25	.37	1.60	.61

[a]The variables are the within-school regression of Science Total Score on Word Knowledge Total score (slope), school means and standard deviations on Total Science and Total Word Knowledge, and school means on sex of student, father's occupation, books in the home, degree to which students report the use of discovery methods in science study, and a composite of student reports of hours of instruction and homework in all science courses. The science total scores used in this analysis have been transformed (10*√Science Total Score) to remove a slight degree of curvilinearity evident in the overall regression of science scores on word knowledge.

Table 2. School-level regressions of means, slopes, and standard errors of estimate on school means on background and schooling characteristics for science achievement of U.S. 14-year olds (N = 107 schools).

	Unstandardized			Standardized		
Explanatory Variables	Mean	Slope	Standard Error of Estimate	Mean	Slope	Standard Error of Estimate
Word Knowledge	.897 (8.07)[a]	.040 (2.25)	-.027 (.40)	.487	.254	-.047
Sex	-5.289 (3.91)	-.427 (1.99)	-.845 (1.06)	-.188	-.179	-.099
Father's Occupation	.583 (2.43)	-.019 (.49)	.086 (.61)	.164	-.063	.079
Books in the Home	3.569 (4.18)	-.058 (.43)	.966 (1.92)	.294	-.057	.262
Instructional Approach	.248 (1.78)	.062 (2.83)	.071 (.86)	.089	.265	.084
Science Instruction	.811 (2.28)	.168 (2.98)	.442 (2.10)	.112	.273	.200
Constant	17.093	-.178	1.566			
R^2	.77	.21	.15			

[a] t statistics in parentheses

Table 3. Predicted science scores for students at various levels of verbal ability from schools with different levels of exposure to science (SCIINST) and emphasis on discovery approach to instruction (EXPLORE).[a]

Level on EXPLORE	Level on SCIINST	Predicted Within-School Slope[b]	Predicted Science Score When Verbal Score = 22	27	32	Difference Between Prediction at Verbal Score of 22 and 32[c]
+1	+1	.97	51.91	56.76	61.61	9.70
+1	0	.87	52.46	56.81	61.16	8.70
0	+1	.87	52.46	56.81	61.16	8.70
0	0	.77	53.07	56.87	60.67	7.60
0	-1	.67	53.57	56.92	60.27	6.70
-1	0	.67	53.57	56.92	60.27	6.70
-1	-1	.57	54.12	56.97	59.82	5.70

[a]The levels are for schools with one standard deviation above the mean (denoted by +1), at the mean (0), and one standard deviation below the mean (-1) on combinations of EXPLORE and SCIINST. For example, a hypothetical school with the combination (+1,+1) has an EXPLORE score of 11.44 and a SCIINST score of 2.20.

[b]These slopes are predicted from the model for the within-class slope in Table 2 when all explanatory variables except EXPLORE and SCIINST have been set at their mean.

[c]The between-student mean and standard deviation of Science Test scores are approximately 57.28 and 8.1 respectively.

Figure 1.

Plots of within-school regressions of science scores on verbal ability for ten lowest and ten highest schools on mean emphasis on discovery methods (EXPLORE).

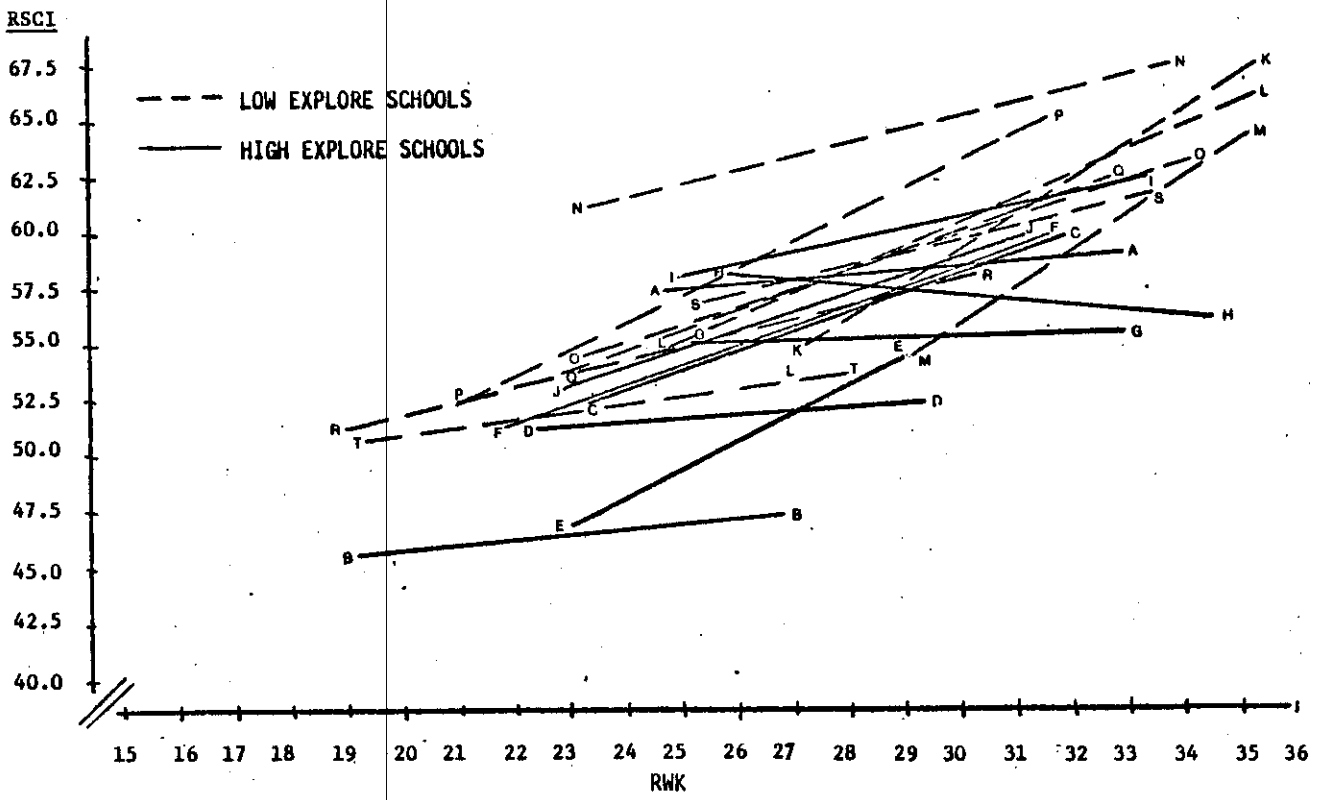
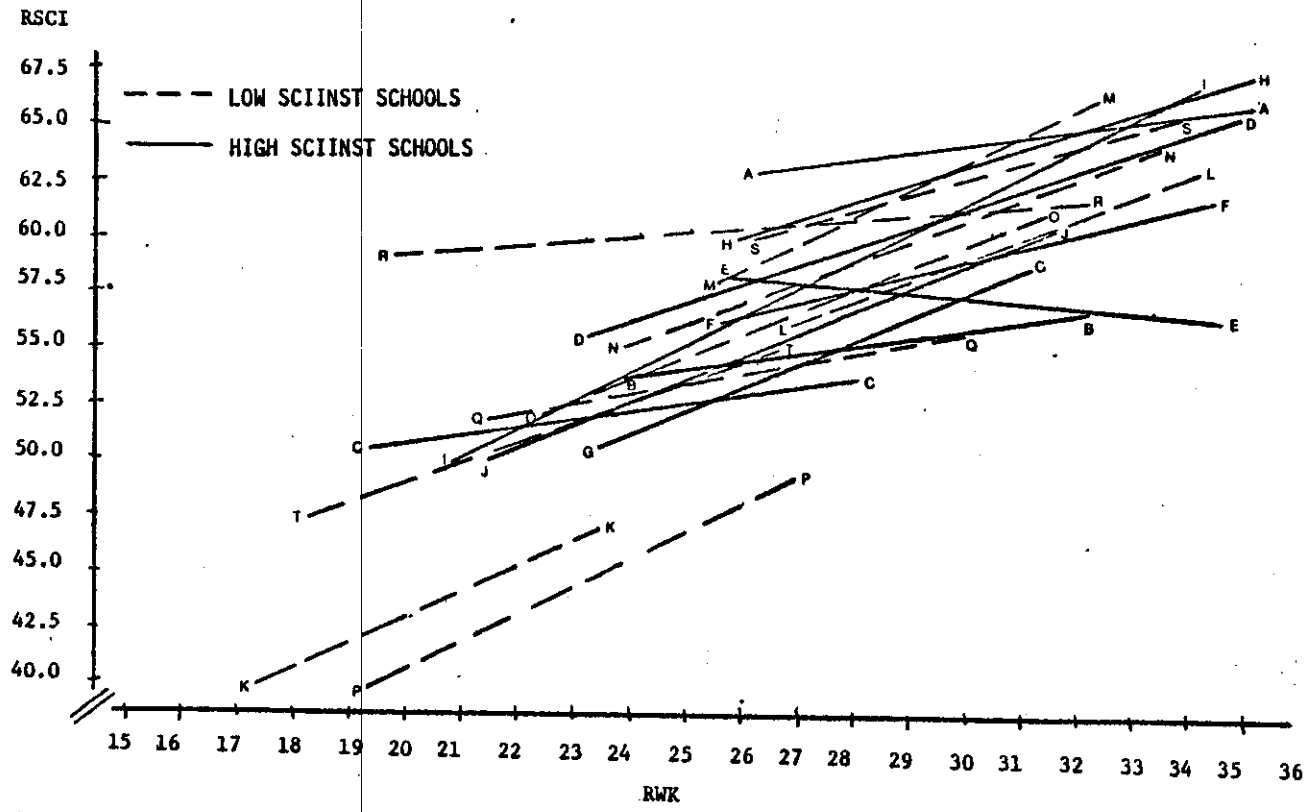

RSCI
67.5
65.0
62.5
60.0
57.5
55.0
52.5
50.0
47.5
45.0
42.5
40.0
LOW EXPLORE SCHOOLS
HIGH EXPLORE SCHOOLS
15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
RWK

Figure 2

Plots of within-school regressions of science scores on verbal ability for ten lowest and ten highest schools on mean exposure to science instruction (SCIINST).


RSCI
67.5
65.0
62.5
60.0
57.5
55.0
52.5
50.0
47.5
45.0
42.5
40.0
LOW SCIINST SCHOOLS
HIGH SCIINST SCHOOLS
15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
RWK