MODELS COMPARING ESTIMATES OF SCHOOL EFFECTIVENESS BASED ON CROSS-SECTIONAL AND LONGITUDINAL DESIGNS

By

MINSUK SHIM

B.A., The Seoul National University, 1982

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY
AND SPECIAL-EDUCATION

We accept this thesis as confirming

to the required standard.

THE UNIVERSITY OF BRITISH COLUMBIA

April 1991

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Educational Psychology and Special Education

The University of British Columbia
Vancouver, Canada

Date March, 15, 1991

Abstract

The primary purpose of this study is to compare the six models (cross-sectional, two-wave, and multiwave, with and without controls) and determine which of the models most appropriately estimates school effects. For a fair and adequate evaluation of school effects, this study considers the following requirements of an appropriate analytical model.

First, a model should have controls for students' background characteristics. Without controlling for the initial differences of students, one may not analyze the between-school differences appropriately, as students are not randomly assigned to schools.

Second, a model should explicitly address individual change and growth rather than status, because students' learning and growth is the primary goal of schooling. In other words, studies should be longitudinal rather than cross-sectional. Most researches, however, have employed cross-sectional models because empirical methods of measuring change have been considered inappropriate and invalid. This study argues that the discussions about measuring change have been unjustifiably restricted to the two-wave model. It supports the idea of a more recent longitudinal approach to the measurement of change. That is, one can estimate the individual growth more accurately using multiwave data.

Third, a model should accommodate the hierarchical characteristics of school data because schooling is a multilevel process. This study employs an Hierarchical

Linear Model (HLM) as a basic methodological tool to analyze the data.

The subjects of the study were 648 elementary students in 26 schools. The scores on three subtests of Canadian Tests of Basic Skills (CTBS) were collected for this grade cohort across three years (grades 5, 6 and 7). The between-school differences were analyzed using the six models previously mentioned. Students' general cognitive ability (CCAT) and gender were employed as the controls for background characteristics.

Schools differed significantly in their average levels of academic achievement at grade 7 across the three subtests of CTBS. Schools also differed significantly in their average rates of growth in mathematics and reading between grades 5 and 7. One interesting finding was that the bias of the unadjusted model against adjusted model for the multiwave design was not as large as that for the cross-sectional design. Because the multiwave model deals with student growth explicitly and growth can be reliably estimated for some subject areas, even without controls for student intake, this study concluded that the multiwave models are a better design to estimate school effects. This study also discusses some practical implications and makes suggestions for further studies of school effects.

Table of Contents

List of Tables

1. Introduction

1.1. Background of the Problem

The evaluation of the effectiveness of school systems has received considerable attention by educators, policy makers and the public (Murnane, 1987). The public continues to demand educational accountability. They want to know whether they are receiving a good return for the money they invest in education, and they call for improved accountability procedures. To meet these demands, educators and policy makers are interested in monitoring student achievement and identifying 'effective' schools—those that enhance student achievement more than other schools. They have systematically collected 'performance indicators' of school effects while examining the following important questions: Which characteristics of teachers and schools are associated with student achievement? To what extent? How can limited resources be allocated most effectively?

Many researchers have examined the relative efficiency of various school policies and management practices on student learning over time. They have asked whether educators should concentrate on certain types of policies or should finance others. Data on performance indicators have been used for evaluation and policy decisions at different levels of the educational system and recently have been used in school award programs in California, Florida, and South Carolina (Mandeville & Anderson, 1987).

Attempts to identify 'effective schools' have created many controversies over

the kind of data to be collected, the appropriate methodology and the interpretation of results. Data on performance indicators should be carefully collected and interpreted for comparing schools. Some school districts and provincial evaluation programs make only simple comparisons of schools' mean scores on some standardized tests. Even though this is the easiest way to compare schools, it is not a fair and adequate evaluation. Through formal and informal selection procedures, schools differ in their student composition. School settings could be considered 'treatments' in a quasi-experimental design. Because random assignment to schools is not possible, every possible variation among schools associated with student background should be controlled to reject other plausible rival hypotheses regarding between-school differences. Otherwise, the school (treatment) effect can be confounded with other unspecified factors (Campbell & Stanley, 1966). Without considering the differences in the characteristics of students when they enter school, such as their general cognitive ability, prior achievement, and family background, it does not make sense to identify schools that are exceptional in their performance. A comparison of school means will falsely suggest that schools with more advantaged students do better than those with less advantaged students.

Moreover, unless all the variables associated with student background are controlled, estimates of school effects will be biased. Students' family background variables alone do not appear to be sufficient for controlling initial differences (Willms, 1985). A premeasure, parallel to the outcome measure of interest, is known to be by far the most important control variable to adjust for student background

(Alexander & Pallas, 1983). Therefore, many researchers have tended to move away from cross-sectional designs toward longitudinal designs (Willms, 1985; Gray, 1988).

However, true longitudinal designs—designs measuring individual change over time—have not been widely used until recently. Researchers who collected data at two or more points in time analyzed them as if they were separate cross-sectional studies. Many authors persuaded empirical researchers not to use the difference score, which is the natural way of measuring change, and to reframe the questions about change into the questions about status (Harris, 1963; Cronbach & Furby, 1970; Linn & Slinde, 1977). More recently, however, a few researchers have argued that the previous discussions about the difference score caused a lot of misunderstandings about the measurement of change. They emphasize the advantages of multiwave longitudinal models over traditional cross-sectional models (Rogosa, Brandt, & Zimowski, 1982; Rogosa & Willet, 1983; Willet, 1988).

The development of appropriate analytical models of school effects is necessary for fair and adequate evaluation of school practices, and for effective decisions about school policies. Such a model should be able to take account of all the relevant factors that affect student achievement. It should provide also more valid estimates of school effects to compare schools. This may be one of the reasons that researchers on teacher and school effectiveness have produced inconsistent findings until now. There have been discrepancies between the models they have used and the inferences they have drawn from the models.

1.2. Definition of Terms

The following terms, with their accompanying definition, will be used throughout this study.

Cross-sectional design: A research design which employs data collected at a single point in time.

Longitudinal design (Growth design): A research design which employs data collected at more than one time point. This design includes both two-wave and multiwave designs.

Two-wave design: A research design based on measures at two time points, such as a Pretest-Posttest Design (Campbell & Stanley, 1966, p.13), or a posttest design with prior measures of ability or academic achievement.

Multiwave design: A research design based on measures at more than two time points, such as a time series design in which same individuals are assessed on three or more occasions.

Difference score: Simple gain score between two time points.

Rate of growth: The average increase in students' scores over a fixed period.

1.3. Identification of the Problem

The question whether schooling has significant effects on student outcomes has attracted researchers for at least three decades. They have tried to assess causal relationships between school characteristics and student achievement using

various models. The review by Bridge, Judd, & Moock (1979) summarizes the works done during the late sixties and seventies. Recently, several review articles have noted the conceptual and methodological problems which hampered the earlier studies of school effects. There have been two major methodological problems: the problem of measuring change (Bryk, 1980; Rogosa & Willet, 1983; Willet, 1988), and the problem of analyzing multilevel data (Burstein, 1980; Aitkin & Longford, 1986; Goldstein, 1987; Raudenbush & Bryk, 1988). There is no doubt that student learning and growth is of central interest to educational researchers and that such growth occurs in hierarchical settings.

Until now, measuring 'change' has not received the attention it deserves. Many influential authors condemned the difference score as unreliable or invalid. Cronbach & Furby (1970) suggested that questions of learning should be framed as questions of 'status' rather than questions of 'growth'. Rogosa and Willet (1983), among others, recently argued that this was simply an inappropriate conceptualization of 'change': it had been conceptualized as an 'increment' in the time period between premeasure and postmeasure rather than as a process of continuous development over time (Willet, 1988, p. 347). This misconceptualization suggested that measuring change was limited exclusively to two-wave designs. But recent longitudinal approaches describe 'change' as continuous growth based on more than two time points. By collecting additional waves of data, individual growth can be measured more appropriately and accurately (Willet, 1988; Bryk & Raudenbush, 1988).

The purpose of the present study is to compare various models for assessing school effects, and to demonstrate the benefits of modelling growth rather than status. The analyses provide estimates of school effects based on three research designs: cross-sectional, two-wave, and multiwave. With each design, I compare a model that includes no adjustment for students' gender and ability with one that includes control for students' gender and prior ability. The study is restricted to the comparisons of academic performance and does not address broader questions about school effectiveness. The study has implications for those collecting indicators of school performance at school, school district, or provincial levels, because it examines how performance data should be collected and interpreted for a fair evaluation of school effects.

1.4. Research Questions

This study assesses differences among schools using the six models listed below:

Ia. Cross-sectional model of grade 7 status without control for students' gender and prior ability.

Ib. Cross-sectional model of grade 7 status with control for students' gender and prior ability.

IIa. Longitudinal model of growth based on measures at two points in time, without control for students' gender and prior ability.

IIb. Longitudinal model of growth based on measures at two points in time, with control for students' gender and prior ability.

IIIa. Longitudinal model of growth based on measures at three points in time, without controls for students' gender and prior ability.

IIIb. Longitudinal model of growth based on measures at three points in time, with controls for students' gender and ability.

In comparing these models, four research questions were addressed.

1. Are there significant differences among schools in their average levels of academic achievement at the end of grade 7, after controlling for students' gender and prior ability? (Model Ia & Model Ib)

2. Are there significant differences among schools in their students' average rates of growth in academic achievement between grades 5 and 7, after controlling for students' gender and prior ability? (Model IIIa & Model IIIb)

3. To what extent do estimates of average rates of growth (adjusted for gender and prior ability) based on a two-wave model differ from those based on a multiwave model? (Models IIa, IIb & Models IIIa, IIIb)

4. To what extent are estimates of average levels of achievement or average rates of growth biased if there is no adjustment for students' gender or prior ability?

The second chapter reviews the major literature related to two basic problems in most educational studies and discusses the procedures developed to resolve these problems. The *Hierarchical Linear Model (HLM)* that was employed as a key

methodological tool of this study is also reviewed. Chapter 3 outlines the sample, the data collected, and the models used in the present study. The fourth and fifth chapters present the results and conclusions from the investigation.

2. Review of Literature

This chapter outlines two questions relevant to the current research questions: Why and how should we measure change? How can we analyze multilevel data that describe many educational processes? These are not separate issues. As Burstein (1980) claimed, the specification of appropriate analytical models can be the answer for both questions. Among several statistical models that have been developed to address these problems, this study reviews the Hierarchical Linear Model (HLM) in detail (Bryk & Raudenbush, 1988; Raudenbush & Bryk, 1986, 1988). The HLM has been used in a variety of educational studies, such as studies of school effectiveness (Raudenbush & Bryk, 1986; Willms & Raudenbush, 1989), and studies of individual growth trajectory or measurement of change (Bryk & Raudenbush, 1987, 1988; Willms & Jacobsen, 1990). This chapter also examines various applications of the HLM.

2.1 Measurement of Change

Research studies attempting to compare schools in terms of their performance indicators must address an important question: Should we use a 'snapshot' of achievement as an indicator of school quality? Or should we use the rate of growth as an indicator of school quality? The review comprises three categories of design: cross-sectional, two-wave, and multiwave designs.

Cross-Sectional Designs

In earlier studies, researchers tried to assess schooling effects at a single point in time with a cross-sectional design. They tried to explain the relationships between certain schooling inputs and student outcomes measured at one time point, after controlling for students' background characteristics. A major concern about the cross-sectional design was related to the problem of 'selection bias'. Unless all of the relevant variables were controlled, the estimates of school effects would be biased. This was one of the major criticisms about the historical report by Coleman, Hoffer, and Kilgore (1981). They analyzed data from the first wave of the High School and Beyond (HSB) study and attempted to estimate differences between public and private schools in their average performance level. They controlled for differences between the two sectors in their background characteristics using student-level variables such as family income, parental education, race, ethnicity, and family structure. But the study was limited in that the most important control variable, a measure of prior achievement or aptitude before high school was not available in the 1980 data. The accuracy of their estimates depended on their statistical model to control for selection bias. Many critics attempted to address this problem (Alexander & Pallas, 1983; Willms, 1985). Therefore, when the 1982 data became available, subsequent analyses of the data were able to estimate the sector effects more accurately using a longitudinal design—see Haertel, James & Levin (1987) for a detailed review.

In addition, longitudinal designs have a conceptual advantage over

cross-sectional designs, because they deal with individual 'growth' rather than 'status'. A cross-sectional design can at best represent the prevailing 'status' measured at one time, whereas the fundamental questions of educational research ought to be concerned with individual learning that implies change and growth (Willet, 1988, p.346).

Two-Wave Designs

The most frequently used longitudinal design has been a two-wave design, a traditional Pretest-Posttest design (Cook & Campbell, 1979). The typical pretest-posttest design consists of an assessment of individual performance using the same measure before and after treatment. If subjects are not randomly assigned to groups, we need to adjust for preexisting differences among groups to assess the valid treatment effect. There are several ways of adjusting for pre-treatment group differences—see Anderson, Auquier, Hauck, Oakes, Vandaele, & Weisberg (1980) for a detailed review. Analysis of covariance (ANCOVA) and the use of difference scores or gain scores are two of the most popular methods of analyzing pre- and postmeasure data. We compute the mean gains across groups and then the treatment effect is nothing but the difference in these mean gains among treatment groups.

Nevertheless, the statistical properties of the difference score have become the focus of criticism in theories of measurement. Many authors condemned the difference score as an unfair measure of individual growth because of its low

reliability, and the fact that it is negatively correlated with initial status (Bereiter, 1963; Linn & Slinde, 1977; Cronbach & Furby, 1970).

First, they argued that the difference score cannot be both reliable and valid simultaneously. Based on the classical test score model, they argued that the difference score cannot be interpreted as a valid estimate of individual growth unless the premeasure and postmeasure were highly correlated. They also argued that this requirement for construct validity makes the difference score intrinsically unreliable according to the following expression of reliability:

$$\rho_D = \frac{\sigma^2_{X_1}\rho_{X_1} + \sigma^2_{X_2}\rho_{X_2} - 2\sigma_{X_1}\sigma_{X_2}\rho_{X_1X_2}}{\sigma^2_{X_1} + \sigma^2_{X_2} - 2\sigma_{X_1}\sigma_{X_2}\rho_{X_1}\rho_{X_2}}$$

Second, in addition to low reliability, many empirical findings have shown that the correlation between the difference score and the pretest (r_{X1D}) is typically negative. Based on this, Linn and Slinde (1977) argued that the difference score is inappropriate to measure individual change, because it gives "an advantage to persons with certain values of the pretest scores" (p.125). Concerned with the negative bias of sample correlation, some authors recommended the use of the residual change score which is uncorrelated with initial status (Cronbach & Furby, 1970). But it is hardly interpretable in many cases and does not provide much different estimates of individual growth from those based on the simple difference score.

Those purported problems of the difference score are basically due to

measurement error and not the problems of the difference score *per se*. The negative bias of r_{X1D} is caused by measurement error in the estimation of correlations. The sample correlation r_{X1D} is not a good estimate of the correlation between true change and true initial status $\rho_{\epsilon 1\beta}$. The only correlation of any real interest is the population correlation between true change and true initial status (Rogosa, Brandt & Zimowski, 1982). This correlation can be estimated if we can separate the true parameter variance from the observed variance. Low reliability of the difference score is also related to measurement error. The important ideas are that reliability is a ratio of true variance to observed variance, and that the difference scores are notoriously unreliable because the observed variance is exceedingly large due to the fact that it includes the variance due to measurement error from both occasions.

The true variance, however, includes the variance due to systematic differences among individuals. The classical test score framework overlooks the possibility of variation in the rates of growth among individuals: some might grow faster or slower than others. Individuals may change at different rates during different time periods. When true differences in the rates of growth exist, the true variance becomes larger, and the ratio of true variance to observed variance (reliability) can be high, even higher than the reliability of the scores themselves. In other words, the difference score is not intrinsically unreliable if there are substantial differences among individuals in their rates of growth.

Even when the estimated reliability is low, the difference score can be an

accurate and useful measure of change, because "low reliability does not necessarily imply lack of precision" (Rogosa *et al.*, 1982, p.731). If individuals grow at the same rate, the reliability of the difference score will be low no matter how precisely the difference score is measured. That is, low reliability comes from systematic variations among individuals as well as measurement error. Rogosa and Willet (1983) have successfully demonstrated the statistical properties of the difference score under different circumstances, and showed that the difference score is an unbiased estimate of individual growth.

Multiwave Designs

The difference score between two data points might be unreliable due to large measurement error and the lack of variation among individual growth. However, the precision of the difference score can be improved with data collected at more than two time points. The precision of the parameter estimates increases as the number of data points increases. Moreover, two-wave data require external information about precision, whereas precision can be estimated with multiwave data. In addition, the two-wave design does not give enough information about individual growth. The model fails to unambiguously provide evidence about the shape of individual growth: whether it is linear or nonlinear over time. With only two waves of data, growth is assumed to be linear. Even if the constant rate of growth model (linear model) appears to be appropriate locally—quite likely in most cases, it is still interesting to see whether the rate of growth might be nonlinear in

the long run.

Hence, Willet (1988) suggested the collection of additional waves of data instead of trying to fix up the flaws of two-wave designs with other fallible statistical controls. With multiwave data, a researcher can more confidently specify an appropriate statistical model for assessing individual growth.

The alteration from a question of 'status' to a question of 'change', and the extension of a simple two-wave longitudinal design to a multiwave design, appear to be quite logical. But unfortunately, multiwave data have been used rarely in the literature of measuring change. Researchers have disregarded the advantages of multiwave data: if the simple difference score is intrinsically unreliable, why should one bother to collect additional data anyway? But such arguments are no longer reasonable when the difference score can be both reliable and valid, and the multiwave design can improve the precision of growth parameters of the two-wave design.

2.2 Analysis of Multilevel Data

Another troublesome and long-standing methodological concern in educational research has been related to the problem of analyzing multilevel data. Schooling is a multilevel process. Students receive schooling in classrooms, classrooms are nested within schools, and schools are nested within school districts. Educational decisions are made at different levels of this hierarchy (Barr & Dreeben, 1983). But

traditional models in educational research have employed single-level analyses. By ignoring the obvious multilevel structure, these single-level models did not match with the nested multilevel educational processes (Haney, 1977). Thus, Burstein (1980) called for the development of an appropriate statistical model that could specify the processes occurring within each level of an hierarchy. Such models should specify how schooling inputs measured at different levels of aggregation influence student outcomes.

Burstein (1980) summarized the following four methodological difficulties of analyzing nested data:

"(1) cross-level inference or aggregation bias,

(2) choice of the appropriate unit of analysis,

(3) contextual analysis, and

(4) specification of appropriate analytical models for multilevel data." (p.161)

Aggregation Bias

Issues of 'cross-level inference' arise when researchers make inferences about individual behavior from analyses based on school-level data. The aggregate analyses have been common in educational research because disaggregated data are usually difficult to obtain and complex to analyze. However, aggregate analyses are likely to provide biased estimates of student-level effects. If the purpose of the study is to evaluate the educational effects on individual student's performance, it would not be appropriate to use aggregate data because the level of aggregation

should match the level at which one wishes to make inferences (Haney, 1977).

Choosing the Appropriate Unit of Analysis

Concerns about 'aggregation bias' have been closely related with the issue of 'choosing the appropriate unit of analysis.' Researchers first noticed the problem of ignoring 'classroom' factors and choosing 'student' as a unit of analysis. If 'treatment' is implemented in a classroom setting, ignoring the classroom factor (grouping factor) may violate the assumption of independent response required for performing analysis of variance. Such violation will lead to liberal tests of significance and increase the probability of Type I error (Aitkin, Anderson & Hinde, 1981). To avoid this problem, some researchers (for example, Wiley, 1970; Glass & Stanley, 1970) suggested that the appropriate unit of analysis is the class where instruction is received simultaneously by all students. On the other hand, the researchers who supported the choice of students as the unit of analysis, asserted that student reaction is individual and thus the focus of the evaluation should be on individual student—see Wittrock & Wiley (1970) for a detailed discussion. Some researchers chose to analyze their data in both ways, that is, at both the student and classroom level. But this strategy provided no guidance on how to proceed if the results diverged, as occurred in the evaluation of Project Follow Through (Haney, 1977).

Burstein (1980), among others, argued that the 'choice of unit' is not the right question. Because relationships between variables at one level influence

relationships at other levels, choosing a single correct unit is simply a digression. The emphasis should be on choosing an appropriate analytical model which allows the estimation of random variation at both levels rather than choosing a specific unit of analysis (Burstein, 1980, p.196).

Contextual Effect

Another consideration for examining school effects is the 'contextual effect' of group membership. The 'contextual effect' is "the effect that a group's aggregate characteristics have on outcomes, over and above the effects due to the individual-level characteristics." (Raudenbush & Willms, 1988)

For example, the school mean of ability has been used to represent the ability context of the school. A contextual effect for ability occurs when school mean ability is related to individual outcomes after controlling for the effect of individual ability and other relevant individual-level factors. Individual ability affects performance. Moreover, school mean ability may affect instructional practices by causing schools to adjust their instruction to the level of the students' ability. As a result, individual students within the school can be expected to learn more or less than they would have in other schools. In school effects research, the contextual effect is viewed as independent of one's own ability whereas the so-called 'frog-pond' effect represents the interaction effect between school mean ability and individual ability. They are not easily separated. Furthermore, the contextual effect includes the influence of wider social, economic and political factors of community that are

usually intractable, such as community support for education, local resources, and residential segregation (Dyer, 1970; Willms & Raudenbush, 1989). That is, the contextual effect includes all the effects that are not directly manipulable by school after controlling for students' background characteristics.

Therefore, they are somewhat difficult to estimate. Research, however, has shown that it is important to examine school effects with and without controls for contextual effects, that is, Type A and Type B effects (Raudenbush & Willms, 1988). And the contextual effects that are not easily manipulable by school should be distinguished from substantive school policies and practices when the purpose is to hold teachers and principals accountable. They are not responsible for the factors which lie beyond their control. With an appropriate model, we should be able to specify all the relevant group-level variables which affect individual outcomes.

Specification of Appropriate Analytical Model for Multilevel Data

All the above problems call for one solution: the specification of an appropriate analytical multilevel model (Raudenbush & Bryk, 1988). An appropriate statistical model should identify and sort out the effects attributable to the different levels within the educational system. If we can specify such a model, all the other problems which threaten the valid inferences of multilevel data can be resolved. Several analytical models have been developed (Dyer, 1970; Dyer, Linn, & Patton, 1969; Marco 1974; Aitkin & Longford, 1986; Goldstein, 1986; Raudenbush & Bryk, 1986). Among them, this review now turns to the recently developed hierarchical

linear model.

2.3. Hierarchical Linear Model

Since the publication of Burstein's (1980) review, several groups of researchers have developed statistical models which can resolve the problems associated with multilevel data. These models have been called variance component models (Aitkin & Longford, 1986), mixed linear models (Goldstein, 1986), and hierarchical linear models (Raudenbush & Bryk, 1986,1987). Despite the differences in algebraic operations, the above methods share common properties. First, they provide explicit structural models for nested processes occurring within each level. These models explicitly specify how the explanatory variables at a higher level of aggregation influence outcomes at a lower level. Second, these methods allow researchers to specify and test the random effects of each unit. That is, a researcher can test whether the slopes of the control variables are homogeneous across schools. Whether the extent of random effects are statistically significant, the variation among subgroups and among schools is of interest. Among these methods this study will adopt the term, 'hierarchical linear model' developed by Raudenbush and Bryk, which will be labelled HLM for convenience.

This review first describes the applications of the HLM for the studies of school effectiveness and individual growth. Then, it describes a three-level HLM, which is a promising approach to the combined problem of school effectiveness on individual growth.

Application of the HLM for Studies of School Effectiveness (Multilevel Data)

The general framework for analyzing school effects is multiple regression, with schooling outcome regressed on variables describing student intake and school characteristics. Although many researchers have noticed mismatches between the single-level linear models and the multilevel educational processes, such traditional linear models have been widely used because there have been no viable alternatives. One of the promising approach to resolve these problems was the 'slopes-as-outcomes' model (Burstein, Miller & Linn, 1979). This approach allows researchers to explore the effects of schooling inputs on the structural relationships within schools (slopes) as well as to examine the effects of schooling inputs on average performance (intercepts). In spite of these advantages, this method has not been widely used because of a number of technical difficulties: precision of the estimated slopes (unreliability of slopes), distinction between parameter and sampling variance, and multiple slopes as outcomes (Raudenbush & Bryk, 1986). Recent advances in statistical theory, such as EM (Estimation and Maximization) algorithm or Bayesian estimation, have contributed to overcome some of these difficulties. As a 'slopes-as-outcomes' model, the HLM utilizes both intercept and slope coefficients as outcome variables at the next stage. It provides a powerful tool that permits a separation of within-school variation from between-school variation, and therefore allows the researcher to distinguish parameter variance from observed variance.

A two-level HLM for analyzing school effects can be specified with two sets of equations. The first set comprises within-school equations which describe the relationships between an outcome measure and students' background variables, such as socioeconomic status, home environment, previous achievement, and level of general ability. The second set comprises between-school equations which describe the relationships between the regression coefficients from the first set (i.e., intercept and slopes) and various school-level variables, including school policy and context variables.

Following the notation of Willms and Raudenbush (1989), the within-school regression model can be written as;

$$Y_{ij} = \beta_{j0} + \beta_{j1}X_{ij1} + \ldots + \beta_{jk}X_{ijk} + \epsilon_{ij} \qquad (1)$$

where Y_{ij} is the outcome score for student i ($i=1,\ldots,n_j$) in school j ($j=1,\ldots,J$). There are k independent variables, X_{ijk}, which describe students' background characteristics. The β_{jk}'s are within-school regression coefficients and the ϵ_{ij} are student-level residuals. If the background variables are centered around their means for the entire sample, the estimates of the intercept, β_{j0}, are the background-adjusted school means. They describe how well a student with sample-average background characteristics can be expected to score in each school. The estimates of slopes, β_{j1}, β_{j2}, ..., β_{jk} describe the effect of each background variable on the outcome score. When Y_{ij}'s and X_{ij}'s are standardized, the coefficients are equivalent to the partial correlations between the outcome and the background variables.

At the second stage, researchers are interested in whether these estimates of intercepts (β_{j0}) and slopes ($\beta_{j1},...,\beta_{jk}$) are a function of particular school policies and practices. The second-level equation can be described as:

$$\beta_{jk} = \theta_{0k} + \theta_{1k}P_j + \theta_{2k}C_j + U_{jk} \qquad (2)$$

where the β_{jk} are the regression coefficients from the within-school equations. Each β_{jk} is regressed on school policy variables, P_j, and on school context variables, C_j. The between-school regression slopes, θ_{1k} and θ_{2k}, capture the effects of school-level variables on the within-school structural relationships (β_{jk}). The between-school residuals, U_{jk}, denote the unique contribution of each school that is not explained by school-level variables in the model. When we do not specify the school-level variables in the model, U_{jk} represent overall school differences. The variation of these residuals is particularly interesting because it represents the extent to which schools vary in their background-adjusted achievement.

Equations (1) and (2) can be combined into a single equation. Only two student-level variables and two school-level variables were considered here.

$$\begin{aligned} Y_{ij} = {} & \theta_{00} && \text{(grand mean)} \\ & + \theta_{01}X_{ij1} + \theta_{02}X_{ij2} && \text{(controls for student background)} \\ & + \theta_{10}P_j + \theta_{20}C_j && \text{(effects of school-level variables)} \\ & + \theta_{11}P_jX_{ij1} + \theta_{12}P_jX_{ij2} + \theta_{21}C_jX_{ij1} + \theta_{22}C_jX_{ij2} && \text{(interaction effects)} \\ & + U_{j1}X_{ij1} + U_{j2}X_{ij2} + U_{j0} && \text{(school-level residuals)} \\ & + \epsilon_{ij} && \text{(student-level residuals)} \qquad (3) \end{aligned}$$

If the student background variables and school characteristic variables are centered around their means for the entire sample, the intercept coefficient, θ_{00} represents the grand mean of the entire sample. The slope coefficients, θ_{01} and θ_{02} denote the effects of student background variables, and θ_{10} and θ_{20} denote the effects of school-level variables. The statistical test for the homogeneity of regression slopes among groups asks whether the variation of β_{0j}, β_{1j}, or β_{2j} is statistically significant. In other words, researchers can test whether the effects of student-level variables differ across schools and if so, whether the variation is associated with particular school-level variables. They might conclude that certain school policies are effective for some schools but not for the others. Moreover, the above equation permits researchers to test whether school-level variables have the same effect on student outcomes across the subgroups with different background characteristics. $\theta_{11}P_jX_{ij1}$, $\theta_{12}P_jX_{ij2}$, $\theta_{21}C_jX_{ij1}$, and $\theta_{22}C_jX_{ij2}$ represent the effects of the interactions between school-level variables and student background variables (e.g., aptitude-treatment interaction). Researchers might conclude that certain school policies are more effective for some students than others within a school. The equation also distinguishes the school-level residuals ($U_{j1}X_{ij1}$, $U_{j2}X_{ij2}$, U_{j0}) from the student-level residuals (ϵ_{ij}).

In equation (3), the distinction between policy variables and context variables was made in the sense that the first are endogenous to school systems but the latter are exogenous. Based on the interests of different groups, the emphasis can be either on the particular effects of school policy or on the overall effects of schooling

including both contextual and policy effects (Raudenbush & Willms, 1988).

However, researchers often do not have measures of school policies and practices (P_j), or data describing contextual factors (C_j). In this case, equation (3) simplifies to become:

$$
\begin{aligned}
Y_{ij} = {} & \theta_{00} && \text{(grand mean)} \\
& + \theta_{01}X_{ij1} + \theta_{02}X_{ij2} && \text{(controls for student background)} \\
& + U_{j1}X_{ij1} + U_{j2}X_{ij2} + U_{j0} && \text{(school effects)} \\
& + \epsilon_{ij} && \text{(student-level residuals)} \qquad (4)
\end{aligned}
$$

Then the school-level residuals represent overall school effects (Type A effects).

Application of the HLM for Measuring Change

An hierarchical linear model can be used for the measurement of change. In a longitudinal HLM, time is 'nested' within each subject analogous to subjects nested within school. Therefore, the first level of the HLM (within-subject) represents the initial status and individual growth rate for each subject. In the second level (between-subject), the growth parameters of each subject become outcome variables and are explained by subject-level background characteristics.

The within-subject equation can be described with a set of regressions that model outcome scores on time:

$$Y_{it} = \pi_{i0} + \pi_{i1}a_{it} + u_{it} \quad (5)$$

where Y_{it} is the outcome score for student i at time t (t=0,1,2,..), and a_{it} is the age of student i at time t. If a_{it} is centered for the first occasion, then π_{i0} describes the initial status of the student i and π_{i1} denotes the rate of growth of student i. The residual, u_{it}, includes both sampling and measurement error, which is the chief advantage of HLM in growth measure applications. An important feature of equation (5) is the assumption that the growth parameters vary among individuals.

The between-subject equation can be formulated to represent this variation. In this equation, the individual growth parameters are a function of students' background variables. The between-subject equation include two sets of equations. One equation examines the relationships between initial status (π_{i0}) and students' background variables: the other examines the relationships between rate of growth (π_{i1}) and students' background variables.

$$\pi_{i0} = \beta_{00} + \beta_{01}X_{i1} + \beta_{02}X_{i2} + \epsilon_{i0} \quad \text{(initial status)} \quad (6a)$$

$$\pi_{i1} = \beta_{10} + \beta_{11}X_{i1} + \beta_{12}X_{i2} + \epsilon_{i1} \quad \text{(rate of growth)} \quad (6b)$$

The HLM estimates how reliably these growth parameters are measured. By separating within-subject variance from between-subject variance, it allows researchers to partition the observed variance in the estimated regression coefficients into two components: sampling variance and parameter variance. Reliability is the ratio of the true parameter variance to the observed variance, analogous to that of classical test score theory. If most of the variability in growth

trajectory were due to sampling error, it is almost impossible to find any systematic relationships between these estimates and the background variables at the second stage. Then one can not detect the significant relations, even if they exist, because the data do not provide enough power. In that case, the percentage of total variance explained may be very small even when the between-subject model is accounting for most of the explainable variance, that is, the parameter variance.

Along with the reliability, the HLM utilizes the empirical Bayes methods for estimating the parameters of interest. The empirical Bayes methods correct the ordinary least square (OLS) growth parameters (slopes) according to their reliabilities. If those slopes are not very reliable, the empirical Bayes estimates are based primarily on the estimated mean slope for the sample. Therefore, outliers resulting from a large sampling error in the OLS slopes are controlled. Empirical Bayes method shrinks these outliers in toward the overall sample mean. As a result, the HLM enables a more accurate measurement of slopes than does OLS.

A Three-level Hierarchical Linear Model

Applications of the HLM for multilevel data and for measuring change can be combined to provide a promising resolution of these two methodological problems (Raudenbush & Bryk, 1988). For a cross-sectional design of school effects, a two-level hierarchical linear model provides a useful tool for conceptualization of the multilevel character of schooling processes. But the cross-sectional design represents the effects of schooling on students' status, which is one instance of

accumulated growth over time. Conceptually, the model which estimates schooling effects on students' growth seems more appropriate than the model which estimates schooling effects on students' status. However, the problem with a two-wave longitudinal design is that this growth rate is measured less reliably. But reliability increases as the number of data points increases. A multiwave design provides a stronger basis for causal inference in nonexperimental studies than a cross-sectional or a two-wave design does (Cook & Campbell, 1979). A two-level hierarchical linear model which was discussed earlier naturally extends to a three-level model as we examine the school effects on individual growth.

A three-level hierarchical linear model can be conceptualized in the following way. The first two levels of the model are identical to those of the two-level model for individual growth. The first level, that is, within-subject level, represents individual growth as a function of individual growth parameters plus random errors:

$$Y_{ijt} = \pi_{ij0} + \pi_{ij1}a_{ijt} + u_{ijt} \qquad (7)$$

where Y_{ijt} is the outcome score for student i in school j at time t.

The second level of the model, between-subject level, represents the relationship between growth parameters (π_{ij0} and π_{ij1}) and students' background characteristics (X_{ij}).

$$\pi_{ij0} = \beta_{00j} + \beta_{01j}X_{ij} + \epsilon_{ij0} \qquad \text{(initial status)} \qquad (8a)$$

$$\pi_{ij1} = \beta_{10j} + \beta_{11j}X_{ij} + \epsilon_{ij1} \qquad \text{(rate of growth)} \qquad (8b)$$

The third level of the model, between-school level, enables specification of how school characteristics influence the distribution of growth within schools.

$$\beta_{00j} = \theta_{000} + \theta_{001}P_j + \theta_{002}C_j + U_{00j}$$ (effects of school-level variables on average initial status for each school) (9a)

$$\beta_{01j} = \theta_{010} + \theta_{011}P_j + \theta_{012}C_j + U_{01j}$$ (effects of school-level variables on the within-school slopes of status and background) (9b)

$$\beta_{10j} = \theta_{100} + \theta_{101}P_j + \theta_{102}C_j + U_{10j}$$ (effects of school-level variables on average rate of growth for each school) (9c)

$$\beta_{11j} = \theta_{110} + \theta_{111}P_j + \theta_{112}C_j + U_{11j}$$ (effects of school-level variables on the within-school slopes of growth and background) (9d)

The empirical Bayes estimation employed in the HLM makes it possible to estimate individual growth trajectory more precisely than does OLS. Therefore, the model is more sensitive to the effects of school-level variables at the next stage. Researchers using a three-level model with multiple time points per student are likely to discover important effects of schools on student growth—effects that could not be discovered with cross-sectional or two-wave data.

3. Research Methodology

3.1 Subjects and Data Collection

The subjects of this study, which were a subsample of the data collected by Willms and Jacobsen (1990), included a large cohort of students enrolled in one school district who had completed their seventh grade during the 1987-1988 school year. There are 32 elementary schools in the district, which serves two cities and surrounding suburban areas. It also includes a large rural, agricultural area. The population is of mixed socio-economic status and includes several racial and ethnic groups.

In the fall of 1988, when the data were collected from the district records, the students of this grade cohort had just begun their eighth grade. The majority of the students were the same age—born in 1974. But approximately 20 percent were one or two years older as they had repeated one or two grades at some time during their elementary schooling. Approximately 2 percent of the students were one or two years younger than their cohort. The entire cohort included 1122 students. In May of each year from 1984 to 1989, while the students were in grades 3 to 7, the district administered the Canadian Tests of Basic Skills (CTBS) and the Canadian Cognitive Abilities Test (CCAT). Among the entire cohort of 1122 students, only the students who had complete CTBS data from grades 5 to 7 were selected. In addition, students were required to have complete CCAT data for grades 3, 4 and 5, because the average CCAT score across three years was computed to denote student's level

of general cognitive ability. Based on these criteria, six schools were excluded from the analysis because they had fewer than ten students who met the selection criteria. The estimates of school effects based on small samples may not be reliable, making it difficult to interpret significant between-school differences. The achieved sample, therefore, included 648 students in 26 schools. The sample size for each school ranged from 12 to 55. A demographic description of the achieved sample is attached in the Appendix.

A concern about the resultant final sample was that it might be biased because low ability students would be more likely to have incomplete test data. Out of 474 students in the cohort who did not meet the selection criteria, CCAT data at grade 7 were available for 207 students. The difference in average CCAT score between the excluded group and the final sample was small even though it was statistically significant ($p<.05$); 103.195 compared with 105.997. The standard deviations of the two groups were almost the same; 12.65 compared with 13.00. Furthermore, the differences between the group means did not vary significantly by schools. Therefore, I do not think that attrition bias would be substantially critical, especially not for the longitudinal models, because loss of subjects with slightly lower ability would have little effect on the average rate of growth for each school.

Outcome measures in this study were the scores of three subtests on CTBS: reading comprehension, vocabulary and the composite scores of mathematics (mathematics has three subtests—computation, concepts, and problem solving). The Canadian Tests of Basic Skills (CTBS) are norm-referenced achievement tests

designed to reflect the continuous nature of skill development in the elementary school (King, 1982). They are more concerned with generalized intellectual skills that are crucial to the educational development rather than separate measures of achievement in content subjects. Grade-equivalent (GE) scores were used to express the continuous development of basic skills, which were necessary for measuring growth. Instead of usual grade equivalent scores, this test employs 'months-of-schooling' as the outcome metric. This is simply a multiple of GE scores. For the 'months-of-schooling' metric, GE scores will be multiplied by 10 (i.e., 59, 69 instead of 5.9, 6.9, respectively) and the variance will be multiplied by 100 (i.e., .12119 instead of .00121). This does not make any differences in interpreting individual growth and gives more accuracy to the estimates because we keep more decimal points.

The most important student-level background variable in this study was the level of general cognitive ability which was measured by the Canadian Cognitive Abilities Tests (CCAT). The CTBS and the CCAT were standardized on the same population of students and administered under the same conditions at approximately the same time. This has made it possible to examine the relationships between achievement and ability under nearly ideal conditions. In addition, the average CCAT scores of verbal, quantitative, and non-verbal subtests across three years were computed to get more reliable estimates of general ability.

Gender was also employed as another control variable for students' background characteristics because prior research suggested significant gender

differences in student achievement (Willms & Jacobsen, 1990).

3.2 Data Analysis

The following models were employed to analyze the data in this study.

Model Ia (Cross-sectional Model without Controls)

This model has been frequently used in school award programs because of its simplicity. But the evaluation purely based on average school performance at a single point in time would not be fair for schools with more disadvantaged students in terms of their background characteristics. If we do not control for the initial differences, the estimates of school effects would be biased. Therefore, for more improved accountability procedure, it might be an interesting question how much differences there are between the estimates of school effects using the model without controls for students' background (Model Ia) and those with controls (Model Ib).

In Model Ia and Ib, my outcome measures were grade 7 CTBS scores in three subject areas (mathematics, reading comprehension, and vocabulary).

The data were analyzed using a two-level hierarchical linear regression model. This model enables us to analyze the multilevel data as discussed in Chapter 2. The data can be described with two sets of equations. The first level of the model simply estimates school mean scores for each school and assumes all other variations are just random errors. The within-school equations are as follows:

$$Y_{7ij} = \beta_{0j} + \epsilon_{ij}$$

where Y_{7ij} is an outcome measure at grade 7 for student i ($i=1,2...,n_j$) who is in school j ($j=1,2,...,26$).

The parameter, β_{0j}, simply represents the average performance of school j, and was employed as an outcome measure in the second level of the model:

$$\beta_{0j} = \theta_{00} + U_{0j}$$

where θ_{00} is a grand mean, that is, an average performance of 26 schools in the sample. Therefore, the between-school residual term, U_{0j}, includes the overall school effects including both context and policy effects of each school and school-level random errors. Unfortunately, no variables about specific school policies and practices were available in this study. Therefore, this analysis examined only Type A effects; that is, overall effects including both school policy and contextual effects (Willms & Raudenbush, 1988). Because the relevant variables at school level were not specified, the estimates of school effects in this analysis can be considered upper limits.

Model IIb (Cross-sectional Model with Controls)

This model has been the most popular statistical model for school evaluation. Since schools do differ in terms of their student composition, this model is designed to control for some of the most important differences of students' background characteristics. Two control measures, CCAT and gender, were employed in this

investigation. No family background variables were controlled in this study. But many researchers argued that measures of general ability are known to be more powerful controls for students' background characteristics than the family background variables (Alexander, Cook & McDill, 1978). In particular, CCAT is a very powerful control on CTBS, since those two tests were standardized on the same population at almost the same time. Gender was another control variable in this analysis because gender differences have appeared to be significant in some content areas (Fennema, 1980; Martin & Hoover, 1987). The first level of the model, therefore, examines the relationships between student outcomes and two background variables:

$$Y_{7ij} = \beta_{0j} + \beta_{1j}(CCAT)_{ij} + \beta_{2j}(Gender)_{ij} + \epsilon_{ij}$$

The intercept parameter, β_{0j}, denotes the expected grade 7 status for a 'typical' child because the background variables were centered such that zero values represent what we refer to a typical child: one with a CCAT score of 100 (the normalized standard mean score for the test).

The slope parameters, β_{1j} and β_{2j} denote the effects of CCAT and gender on student outcomes, respectively. All three parameters were employed as outcomes in the next level of the model. The between-school level equations are as follows:

$$\beta_{0j} = \theta_{00} + U_{0j} \qquad \text{(intercept)}$$

$$\beta_{1j} = \theta_{01} + U_{1j} \qquad \text{(slope for CCAT)}$$

$$\beta_{2j} = \theta_{02} + U_{2j} \qquad \text{(slope for gender)}$$

where the residual terms, U_{0j}, U_{1j} and U_{2j} represent the overall school effects on student outcomes because the school-level variables are not specified in this model. The variation of these residuals is of interest to see whether there were significant differences among schools. The variances of the school-level residuals were tested by a chi-square test of variance. In addition, a researcher can test whether the regression slopes for CCAT or gender are parallel across schools. This is one of the interesting questions a researcher might have but could not ask with the ordinary analysis of covariance method. If there are significant differences in slope coefficients among schools, the HLM allows them to vary; if not, the slopes can be fixed across schools as with OLS.

Model IIa (Two-wave Longitudinal Model without Controls)

The difference scores (gain scores or change scores) have had a long history in educational research because questions in education are naturally concerned with students' gain and change over time. Until recently, a two-wave model has been used most frequently for measuring individual change.

Model IIa is almost identical with Model Ia except that the student-level outcome measures are the difference scores between grades 6 and 7 instead of grade 7 scores. Individual difference scores were partitioned into two parts: average difference scores for schools and individual random errors, assuming the deviations from the average school gain scores represent only the random fluctuation among individuals.

The within-school level equation is followed as:

$$Y_{67ij} = \beta_{0j} + \epsilon_{ij}$$

where Y_{67ij} is the difference score between grades 6 and 7 for student i ($i=1,2,...,n_j$) in school j ($j=1,2,...,26$). Then the parameter, β_{0j}, denotes the school average difference score.

The second level of the model is identical with that of Model Ia.

$$\beta_{0j} = \theta_{00} + U_{0j}$$

The between-school equation estimates the grand mean, which is the average difference score of the sample. Then the school-level residual, U_{0j}, represents everything other than the average difference score of the sample, which includes both systematic differences among schools (overall school effects) and random errors. Again, the chi-square test of variance was used to test the variation of school-level residuals.

The HLM also provides reliability estimates—the ratio of the parameter variance to the observed variance—by differentiating between-school variance from within-school variance. The reliabilities of the two-wave models compared with the cross-sectional models and the three-wave models are of interest in this study.

Model IIb (Two-wave Longitudinal Model with Controls)

This model is similar to Model Ib because it also controls for differences in

students' background characteristics. But the key difference between the two is that the difference score is used as a student outcome in this model.

The first level of the model is as follows:

$$Y_{67ij} = \beta_{0j} + \beta_{1j}(CCAT)_{ij} + \beta_{2j}(Gender)_{ij} + \epsilon_{ij}$$

where Y_{67ij} is the difference score between grades 6 and 7 of student i in school j and the slope parameters, β_{1j} and β_{2j} represent the effects of CCAT and gender, respectively. The intercept parameter, β_{0j} represents the expected difference score between grades 6 and 7 for a typical child with a CCAT score of 100.

The second levels of the model are identical with those of Model Ib.

$$\beta_{0j} = \theta_{00} + U_{0j}$$

$$\beta_{1j} = \theta_{01} + U_{1j}$$

$$\beta_{2j} = \theta_{02} + U_{2j}$$

As in the Model Ib, the residual terms, U_{0j}, U_{1j}, and U_{2j} represent the overall school effects on student outcomes. The variances of them were statistically tested. The slope coefficients, θ_{01} and θ_{02} were also tested to see whether the effects of CCAT and gender are the same across schools.

Model IIIa (Three-wave Longitudinal Model without Controls)

In this model, three waves of data which is the simplest form of multiwave design were employed.

Ideally, one can employ a three-level hierarchical linear model as discussed in Chapter 2: that is, between temporal points within subjects, between subjects within schools, and between schools. Unfortunately, computer programs which can estimate three-level hierarchical linear models have only recently been fully developed. The alternative way to handle this is to use the OLS method to estimate the parameters of the first level. Then using the growth parameters obtained by the OLS, the usual two-level hierarchical linear model can be employed for the two higher levels (Willms & Jacobsen, 1990). The problem with this method might be the precision of individual growth parameters. However, the growth estimates were based on multiple time points. If more and more time points were added, the precision of the growth parameters estimated by the OLS and by the HLM would be quite close. I compared the estimates of growth rate using three data points and those using four or more data points whenever they were available, and found the estimates were quite stable. Therefore I expect the results do not significantly differ from those using a three-level hierarchical model.

The first level of the model can be described as:

$$Y_{ijt} = \pi_{ij0} + \pi_{ij1}(\text{Time})_{it} + u_{ijt}$$

where Y_{ijt} is a growth measure for student i in school j at time t (t=1,2,3). The variable, $(\text{Time})_{it}$, denotes the number of months of schooling that the student i in school j had received before testing occasion t. In this study, $(\text{Time})_{it}$ was set to be zero for the second testing occasion (grade 6) since the estimates of expected values

near the center of the data would be more reliable than those at the extremes with OLS regression.

The intercept parameter, π_{ij0}, represents the status for each student at the end of grade 6, which in this study refers to the GE score attained after 69 months of schooling (including kindergarten). The slope parameter, π_{ij1}, represents the individual rate of growth over three years. Both of the parameters can be used as outcome measures at the second level. However, this study did not employ the grade 6 status as an outcome measure at the second level because the relationships between the grade 6 status and the background variables were not of any practical interest in this study. The rates of growth also allow me to compare the results with those from the previous two-wave models.

The second (between-subject) level of the model can be described as:

$$\pi_{ij1} = \beta_{0j} + \epsilon_{ij}$$

where β_{0j} denotes the average rate of growth for school j and ϵ_{ij} represents the random fluctuation among students.

The third (between-school) level of the model becomes:

$$\beta_{0j} = \theta_{00} + U_{0j}$$

where θ_{00} denotes the grand mean, the average rate of growth for the sample. The between-school residual term, U_{0j}, includes the overall school effects as well as the random fluctuation among schools. An hierarchical linear model allows one to

partition the error variance into three parts (within subject, within school and between school) and provides more precise estimates of growth parameters.

Model IIIb (Three-wave Longitudinal Model with Controls)

The first level of the model is identical with that of Model IIIa to estimate the grade 6 status and the rate of growth:

$$Y_{ijt} = \pi_{ij0} + \pi_{ij1}(\text{Time})_{it} + u_{ijt}$$

The second level of the model investigates the relationships between individual rate of growth and the background variables:

$$\pi_{ij1} = \beta_{0j} + \beta_{1j}(\text{CCAT})_{ij} + \beta_{2j}(\text{Gender})_{ij} + \epsilon_{ij}$$

where β_{0j} represents the average rate of growth for a typical child, and β_{1j} and β_{2j} represent the slope parameters for CCAT and gender, respectively.

Both intercept and slope parameters of the second-level become the school-level outcomes at the next level:

$$\beta_{0j} = \theta_{00} + U_{0j}$$
$$\beta_{1j} = \theta_{10} + U_{1j}$$
$$\beta_{2j} = \theta_{20} + U_{2j}$$

The school-level residual terms, U_{0j}, U_{1j}, and U_{2j} are of primary interest to examine between-school differences in their average rate of growth. The slope

parameters, θ_{10} and θ_{20}, were examined to see whether the regression slopes of three background variables were parallel across schools.

Estimation of Bias

The difference between estimates of school mean performance based on an unadjusted model and those based on an adjusted model was computed to denote the estimates of bias between two models.

The bias can be written as follows:

$$\text{Bias} = S^*_j - S_j$$

where S^*_j denotes the estimates of school average performance from the unadjusted model, and S_j denotes the estimates of school average performance from the adjusted model. Then the mean absolute bias between two models was used to indicate the extent of overall bias of the unadjusted model against adjusted model:

$$\text{Mean absolute bias} = \Sigma \mid S^*_j - Sj \mid /26$$

I examined the extent of bias of Model Ia, IIa, and IIIa against Model Ib, IIb, and IIIb, respectively. Furthermore, the biases of the two-wave models (Model IIa and IIb) against the multiwave model (Model IIIb) were also of interest.

4. Results

This chapter reports the findings generated by this study. The study compares the estimates of school effects based on six models (two cross-sectional, two two-wave, and two multiwave models) across three CTBS subtests (mathematics, reading comprehension, and vocabulary, respectively). Tables Ia through IIIb show the results of the HLM for each model. The chapter also examines the between-school differences in average grade 7 status and in average rates of growth, and it examines the effect of two control variables (CCAT and gender) and the reliability estimates of school means. Table IV shows whether the regression slopes of two control variables are homogeneous across schools. Table V describes the extent of bias between adjusted models and unadjusted models across three subject areas.

School Differences in Grade 7 Status

Table Ia and Ib summarize the HLM results of the cross-sectional model with and without controls for student intake (Model Ia vs. Model Ib) for mathematics, reading, and vocabulary. The first row of Tables Ia and Ib show the average estimates of grade 7 status for three subtests. The unadjusted estimates of grade 7 status (Table Ia) were higher or close to the expected score, which is 79 in terms of months-of-schooling, across all three subtests. The adjusted estimates of grade 7 status (Table Ib), however, were below the expected score for all three subtests.

Table Ia

HLM Results for Cross-sectional Model Without Controls (Model Ia)

Estimated Parameters	Mathematics		Reading		Vocabulary	
(Fixed Effects)	Effect	(S.E.)	Effect	(S.E.)	Effect	(S.E.)
Constant	80.761**	(.777)	77.514**	(.708)	78.224**	(.645)
Estimated Variance Components						
(Random Effects)	Estimate	$(\chi^2)_{25}$	Estimate	$(\chi^2)_{25}$	Estimate	$(\chi^2)_{25}$
Observed Variance	15.979		13.425		11.211	
Parameter Variance	10.610**	(78.693)	7.559**	(60.119)	5.673**	(53.921)
Reliability	.664		.563		.506	

Notes: * $p<.05$
** $p<.01$

Table Ib

HLM Results for Cross-sectional Model With Controls (Model Ib)

Estimated Parameters	Mathematics		Reading		Vocabulary	
(Fixed Effects)	Effect	(S.E.)	Effect	(S.E.)	Effect	(S.E.)
Constant	77.809**	(.641)	74.639**	(.490)	75.670**	(.413)
CCAT	.588*	(.024)	.573*	(.026)	.529**	(.026)
Gender	-.095	(.605)	.382	(.671)	-2.414**	(.674)
Estimated Variance Components						
(Random Effects)	Estimate	$(\chi^2)_{25}$	Estimate	$(\chi^2)_{25}$	Estimate	$(\chi^2)_{25}$
Observed Variance	10.401		6.041		4.363	
Parameter Variance	7.640**	(275.83)	2.622**	(195.93)	.890**	(156.14)
Reliability	.735		.434		.204	

Notes: * $p<.05$
** $p<.01$

The adjusted estimates were lower than the unadjusted estimates because the average CCAT of this sample (105.336) was higher than the national norms of 100. The constant represents the expected score for a 'typical' student in terms of months of schooling. When we control for the higher CCAT of this sample, a 'typical' student of this sample had GE scores that were around one month below the expected score for mathematics (77.8 compared with 79), four months below the expected score for reading comprehension (74.6 compared with 79) and three months below the expected score for vocabulary (75.7 compared with 79).

The next two lines of Table Ib show the effects of two background variables for Model Ib. The level of prior ability (CCAT) was significantly related to grade 7 status across all three subtests. The effect sizes of CCAT on three outcomes were almost identical. Gender differences were negligible at this grade in both mathematics and reading; however, males outperformed females by over two months of schooling on the vocabulary subtest.

The results for mathematics are consistent with the findings from previous research on gender differences. In reviewing studies of mathematical achievement and gender differences, Fennema (1980) reported that no significant differences appeared consistently between males and females during the elementary years of schooling. Martin and Hoover (1987) also reported that there was no gender difference in the composite scores of ITBS mathematics from grades 3 to 8, while females did somewhat better in computation and males showed a better understanding of mathematical concepts. Even though the gender differences are

relatively small during elementary education, researchers generally agreed that the gender gap in mathematics becomes larger during secondary education (Maccoby & Jacklin, 1974; Fennema, 1980; Willms & Jacobsen, 1990).

Previous research on reading comprehension has also shown a slight but consistent advantage of females in general reading ability (Martin and Hoover, 1987; Maccoby & Jacklin, 1974). The gender difference derived in this study was small and not statistically significant.

Vocabulary was interesting, quite different from other subtests. Females had a significantly lower mean on vocabulary at the end of grade 7. Generally, female superiority on verbal tasks at the earlier stage of development has been one of the solidly established generalizations in the field of gender differences. Some of the research has found no consistent gender differences, but whenever the difference was found, it was usually females who obtained higher scores (reviews by Maccoby and Jacklin, 1974). However, these differences favoring girls seem to be reversed during the later years of primary schooling. Martin and Hoover (1987) reported very similar results to this study. Instead of the consistent advantage for males or females that was found on the other subtests, they found a cross-over in the results for vocabulary. In grades 3 and 4, females had a slightly higher mean than males. But in the later grades, males had the significantly higher mean, and the improvement in performance by males was consistent.

The bottom half of the tables show the estimates of observed variance and parameter variance, and reliability of estimates of school means. The parameter

variance represents the variation among schools with the variance due to measurement and sampling errors removed. This is similar to the true score variance of classical test score theory. The chi-square tests showed that parameter variance was statistically significant ($p<.01$) across all subtests. In other words, there were significant differences between schools in their grade 7 status, even after controlling for the effects of students' background variables.

The last row of the tables show the reliability coefficients of intercept parameters. The primary purpose of this study was to examine the between-school differences in the average performance (intercept) rather than the between-school differences in structural relationships (slope). Therefore, only the reliability coefficients of intercept parameters were examined here. The between-school differences were most reliably estimated for mathematics and least reliably estimated for vocabulary: .664 and .735 for mathematics, .563 and .434 for reading, and .506 and .204 for vocabulary, without and with controls, respectively.

School Effects on the Rate of Growth

The second research question (see Chapter 1) is concerned with students' rates of growth between grades 5 and 7 rather than their grade 7 status. The first line of Table IIa and IIb present the estimates of average rate of growth during the last two years of elementary schooling, with and without controls for student background for the three subtests. The unadjusted and the adjusted two-wave models (Model IIa and Model IIb) suggested that the average rate of growth of this

Table IIa

HLM Results for Two-wave Model Without Controls (Model IIa)

Estimated Parameters	Mathematics		Reading		Vocabulary	
(Fixed Effects)	Effect	(S.E.)	Effect	(S.E.)	Effect	(S.E.)
Constant	8.516**	(.387)	8.064**	(.374)	6.514**	(.315)
Estimated Variance Components (Random Effects)	Estimate	$(\chi^2)_{25}$	Estimate	$(\chi^2)_{25}$	Estimate	$(\chi^2)_{25}$
Observed Variance	3.997		4.001		2.852	
Parameter Variance	2.230**	(63.558)	.952	(34.102)	.636	(34.072)
Reliability	.558		.238		.223	

Notes: * $p<.05$
** $p<.01$

Table IIb

HLM Results for Two-wave Model With Controls (Model IIb)

Estimated Parameters	Mathematics		Reading		Vocabulary	
(Fixed Effects)	Effect	(S.E.)	Effect	(S.E.)	Effect	(S.E.)
Constant	8.519**	(.398)	7.758**	(.394)	6.312**	(.337)
CCAT	.070*	(.019)	.063*	(.025)	.047*	(.021)
Gender	.132	(.478)	-.952	(.629)	-1.408*	(.534)
Estimated Variance Components						
(Random Effects)	Estimate	$(\chi^2)_{25}$	Estimate	$(\chi^2)_{25}$	Estimate	$(\chi^2)_{25}$
Observed Variance	3.997		3.943		2.897	
Parameter Variance	2.264**	(65.492)	.934	(35.390)	.706	(34.900)
Reliability	.567		.236		.245	

Notes: * $p<.05$
** $p<.01$

sample was lower than expectation, that is, less than ten months-of-schooling. The average rate of growth for vocabulary was approximately three and a half months below the norm for both Models IIa and IIb (6.5 and 6.3, respectively, compared with 10). The CCAT was significantly related to the rates of growth across three subtests. Gender differences were not significant in mathematics and reading, but males outgrow females by one and a half months of schooling in the vocabulary test. The effects of adjustment of the two-wave models were not as large as those of the cross-sectional models. The chi-square tests suggested that the between-school differences in their rates of growth between grades 6 and 7 were statistically significant for mathematics but not for reading and vocabulary subtests. Because of the large measurement errors related to the two-wave data, the reliability of estimates of average gain were lower than that of the cross-sectional models: .56 for mathematics, .24 for reading, and .25 for vocabulary.

Tables IIIa and IIIb present the HLM results of multiwave models with and without controls (Model IIIa vs. Model IIIb). The first row of Tables IIIa and IIIb shows the estimates of average rate of growth between grades 5 and 7. For this particular sample, students seem to be levelling off in vocabulary and reading: that is, the growth rate between grades 5 and 7 was much less than ten (8.9 and 9.3, respectively), whereas in mathematics it was close to ten (9.9).

One of the most interesting findings from this analysis was that the effects of student background variables were no longer significant when we employed growth rates based on multiwave data as the outcome measure. Even the CCAT did

Table IIIa

HLM Results for Multiwave Model Without Controls (Model IIIa)

Estimated Parameters	Mathematics		Reading		Vocabulary	
(Fixed Effects)	Effect	(S.E.)	Effect	(S.E.)	Effect	(S.E.)
Constant	9.896**	(.306)	9.374**	(.217)	8.885**	(.171)
Estimated Variance Components						
(Random Effects)	Estimate	$(\chi^2)_{25}$	Estimate	$(\chi^2)_{25}$	Estimate	$(\chi^2)_{25}$
Observed Variance	2.459		1.296		.862	
Parameter Variance	1.874**	(116.97)	.552**	(46.654)	.121	(30.649)
Reliability	.762		.426		.140	

Notes: * $p<.05$
** $p<.01$

Table IIIb

HLM Results for Multiwave Model With Controls (Model IIIb)

Estimated Parameters	Mathematics		Reading		Vocabulary	
(Fixed Effects)	Effect	(S.E.)	Effect	(S.E.)	Effect	(S.E.)
Constant	9.845**	(.310)	9.295**	(.227)	8.923**	(.182)
CCAT	.011	(.011)	.016	(.012)	-.002	(.012)
Gender	-.173	(.279)	-.166	(.313)	-1.108*	(.309)
Estimated Variance Components						
(Random Effects)	Estimate	$(\chi^2)_{25}$	Estimate	$(\chi^2)_{25}$	Estimate	$(\chi^2)_{25}$
Observed Variance	2.448		1.311		.853	
Parameter Variance	1.866**	(117.26)	.562**	(47.315)	.121	(31.225)
Reliability	.761		.431		.142	

Notes: * $p<.05$
** $p<.01$

not have significant effects on students' rates of growth across the three subtests. Considering the highly significant effects of CCAT in the cross-sectional models, this finding has practical implications. Even when the data on students' background characteristics are not available, the growth measure based on multiwave data can be a better indicator of school performance, even without controls for student intake. 'Selection bias', which occurs when we do not specify all the relevant variables in the model is no longer a big problem for a multiwave longitudinal model as for a cross-sectional model.

However, significant gender differences in vocabulary growth rate (Table IIIb) were noteworthy ($p<.05$). Female growth rate in vocabulary was significantly smaller than the male growth rate during the last three years of elementary schooling. Since boys are supposed to catch up in vocabulary during the later grades of elementary schooling, the average growth rate of boys between grades 5 and 7 may be significantly larger than that of girls. One suggestion might be that a non-linear model should be used when modelling vocabulary growth rate between grades 3 to 7.

The test results for the parameter variances among schools are shown in the bottom portion of the tables. As we move towards the longitudinal models, the parameter variance gets much smaller because growth measures indirectly control for differences in prior achievement, that is, grade 5 and 6 scores. For mathematics and reading, there were significant differences among schools in their effects on growth rate between grades 5 and 7. The parameter variance of mathematics was

much larger than that of reading. This suggests that mathematics is more related to school instruction than is reading. The parameter variance among schools for vocabulary was smallest and was not statistically significant at the .05 level of significance. This is consistent with what we would expect given home influences on students' vocabulary. Vocabulary is not so directly related to instruction given at school as are the other skills. It depends more on one's language experiences in his or her home. It is possible that schools do not differ significantly in their effects on students' vocabulary over a period of three years, even though the richness of language experiences in a school program may affect students' vocabulary skill over a longer period.

Another interesting finding of this study comes from the reliability of the average growth measure based on a multiwave model. The reliability of the growth model (Model IIIb) was as high as that of the cross-sectional model (Model Ib): .761 compared with .735 for mathematics, .431 compared with .434 for reading, and .142 compared with .204 for vocabulary. The between-school differences in their rates of growth based on the multiwave model (Model IIIb) were more reliably estimated than those based on the two-wave model (Model IIb) for mathematics and reading: .761 compared with .567 for mathematics, .431 compared with .236 for reading.

Differences between Two-wave Model and Multiwave Model

The two-wave model consistently underestimated the average rate of growth across three subtests: 8.159 compared with 9.295 for mathematics, 7.758 compared

with 9.295 for reading, and 6.312 compared with 8.923 for vocabulary. Moreover, the standard errors of these estimates were much larger for the two-wave model than for the multiwave model: .398 compared with .310 for mathematics, .394 compared with .228 for reading, and .337 compared with .182 for vocabulary. In other words, the average growth rate was less precisely estimated and negatively biased for the two-wave model. The effects of CCAT were weak for the multiwave model compared with the two-wave model. The gender differences in vocabulary were statistically significant at the .05 level of significance. In other words, males grew faster than females between grades 5 and 7 in the vocabulary subtest.

Compared with mathematics and reading, the reliability estimate of the growth parameters for vocabulary was disappointingly low (.142). For mathematics, the between-school differences in their rates of growth were reliably estimated using the multiwave model, even more reliably estimated than the between-school differences in their grade 7 status using the cross-sectional model. For vocabulary, however, the reliability of the estimates of between-school differences in their growth rates was low, even lower than that based on the two-wave model. One of the factors that might affect reliability is the variability among schools. For vocabulary, growth rate would not greatly vary among schools after the background variables were controlled. Schools do not have explicit effects on students' vocabulary power as they have in the other two tests. The small variability in the vocabulary growth rates among schools might explain why between-school differences in their vocabulary growth rate were most unreliably estimated.

Homogeneity of Regression Slopes

The analysis also determined whether the effects of background variables on student outcomes varied significantly across schools. For example, some schools might be particularly effective for males but not for females, or *vice versa*, or some schools might be more effective for the students with higher cognitive ability. This is one of the questions that can be determined with an HLM analysis.

Therefore, in a preliminary HLM analysis, all the slopes were allowed to be random. Nine models—three (Model Ib, IIb, and IIIb) for each subtest—were examined. Table IV shows that the variances among schools in their effects of prior ability and gender were not statistically significant at the .01 level of significance. It was not possible to say that some schools were more effective for males than for females, or for students with higher ability than for students with lower ability. Therefore, all the slopes were fixed to be parallel in the main analyses of this study. When slopes are fixed, the intercepts can be estimated more reliably.

Bias of Models

The first part of Table V shows the extent of bias of unadjusted cross-sectional models (Model Ia) against adjusted cross-sectional models (Model IIb) across the three subtests. The extent of bias ranged from -.2391 to 6.5342 for mathematics, from -.0624 to 6.6879 for reading, and from -.6271 to 6.6692 for vocabulary. The mean absolute bias ranged from 2.6002 (vocabulary) to 2.9713

Table IV

Homogeneity of Regression Slopes

subtests	model	slopes	parameter variance	chi-square	p
Mathematics	Model Ib	CCAT	.00713	40.647	.025
		gender	1.5693	23.104	>.5
	Model IIb	CCAT	.00227	28.943	.266
		gender	4.06588	42.960	.014
	Model IIIb	CCAT	.00092	32.037	.157
		gender	.3873	20.286	>.5
Reading	Model Ib	CCAT	.00102	22.872	>.5
		gender	1.79341	28.121	.302
	Model IIb	CCAT	.00375	33.039	.130
		gender	1.33986	19.867	>.5
	Model IIIb	CCAT	.00077	24.225	>.5
		gender	.06767	12.215	>.5
Vocabulary	Model Ib	CCAT	.00189	16.270	>.5
		gender	6.37617	36.098	.07
	Model IIb	CCAT	.00664	35.834	.074
		gender	1.76095	24.370	>.5
	Model IIIb	CCAT	.00108	25.765	.420
		gender	.50950	25.991	.408

Table V

Extent of Bias

	Mathematics	Reading	Vocabulary
Model Ia vs Ib			
min bias	-.2391	-.0624	-.6271
max bias	6.5342	6.6879	6.6692
mean absolute bias	2.9713	2.8802	2.6020
Model IIa vs IIb			
min bias	.051	.1961	.0999
max bias	.7557	.459	.3099
mean absolute bias	.3566	.3059	.2025
Model IIIa vs IIIb			
min bias	-.0144	.0259	-.074
max bias	.1456	.1434	-.0051
mean absolute bias	.0531	.0783	.0374
Model IIa vs IIIb			
min bias	-4.3068	-2.9839	-3.2154
max bias	1.2205	-.0968	-1.9182
mean absolute bias	1.5248	1.2314	2.4087
Model IIb vs IIIb			
min bias	-4.5459	-3.2492	-3.4617
max bias	.8182	-.4675	-2.0573
mean absolute bias	1.7608	1.5373	2.6112

(mathematics). Therefore, estimates of school mean performance based on unadjusted models differ from those based on adjusted models, on average, by about three months of schooling. Moreover, estimates based on an unadjusted model are biased in favor of schools with higher ability students. For example, the school with the highest mean CCAT (115.509) had the highest unadjusted means across all outcomes (91.947 for mathematics, 88.000 for reading and 88.632 for vocabulary). When differences in students' background characteristics were controlled, however, the school with lower mean CCAT (109.231) had higher adjusted means, even though the simple means were much lower than the school with the highest mean CCAT (79.423 for mathematics , 77.346 for reading, and 79.115 for vocabulary). Therefore, mean differences among schools based on the unadjusted cross-sectional model may not be appropriate for comparing schools. A simple comparison of school means will falsely suggest that schools with advantaged students do better than those with less advantaged students.

The second part of Table V shows the extent of bias of unadjusted two-wave models (Model IIa) against adjusted two-wave models (Model IIb). The extent of bias ranged from .051 to .7557 for mathematics, from .1961 to .459 for reading, and from .0999 to .3099 for vocabulary. The mean absolute bias ranged from .2025 for vocabulary to .3566 for mathematics (i.e., less than one-half of one month of schooling). Thus, the bias was far less than that of the unadjusted cross-sectional model. The bias between the unadjusted two-wave model and the adjusted two-wave model was not as crucial as was the bias between two cross-sectional models.

The third part of Table V shows the extent of bias of unadjusted multiwave models (Model IIIa) against adjusted multiwave models (Model IIIb). The extent of bias ranged from -.0144 to .1456 for mathematics, from .0259 to .1434 for reading, and from -.074 to -.0051 for vocabulary. The mean absolute bias was the smallest amongst three designs (.0531 for mathematics, .0783 for reading, and .0374 for vocabulary). This also supports the result that the effect of students' background characteristics was not significant for the multiwave models. This suggests that one could use the rate of growth as a performance indicator of school quality and get fairly accurate estimates of school effects without controlling for students' background characteristics.

The last parts of Table V examine the extent of bias among four longitudinal models. The adjusted multiwave model (Model IIIb) was assumed to be the best longitudinal model amongst the four models. The extent of bias ranged from -4.5459 to 1.2205 for mathematics, from -3.2492 to -.0968 for reading and from -3.4617 to -1.9182 for vocabulary. The mean absolute bias of unadjusted two-wave model (Model IIa) against Model IIIb ranged from 1.2314 to 2.4087. The mean absolute bias of the adjusted two-wave model (Model IIb) against the adjusted multiwave model (Model IIIb) ranged from 1.5373 to 2.6112. These differences were substantially large compared with the bias between two-wave models and the bias between multiwave models. This shows that the two-wave models provide fairly biased estimates of students' growth compared with the multiwave models. The two-wave model is not an appropriate model to estimate students' rate of growth.

5. Summary and Conclusion

This study examined several statistical models designed to detect school differences in their effects on student outcomes. Almost all research studies on school effectiveness has been concerned with school effects on students' average performance or variability at one particular time. That is, studies have been cross-sectional rather than longitudinal. Even though the goal is to assess student learning, which implies change and growth (Willet, 1988), most empirical methods of measuring change have been considered inappropriate and invalid (Harris, 1963; Cronbach & Furby, 1970). Therefore, questions which should be addressed under a longitudinal framework were analyzed with cross-sectional designs.

This study supports the idea of a more recent longitudinal approach to the measurement of change. It claims that not only the idea of measuring change is important, but also the empirical methods of measuring change can be improved by collecting additional waves of data. The discussions about measuring change have been unjustifiably restricted to the two-wave design. But the two-wave model is simply not a very good design of measuring individual growth. We can get better estimates of individual growth with multiwave data.

Therefore, the primary purpose of this study was to examine how much the multiwave models could improve the estimates of individual growth over two-wave models. To accommodate the hierarchical characteristics of educational processes and get more precise estimates of growth, the recently developed hierarchical linear

regression model (HLM) was employed in this study.

The major findings of this study are as follows:

(1) Across the three subtests of CTBS, schools differed significantly in their average levels of academic achievement at the end of grade 7, after controlling for students' prior ability and gender. The effects of students' prior ability were significant across all outcomes but gender effects were not significant for mathematics and reading comprehension scores.

(2) There were significant differences among schools in their average rates of growth in academic achievement between grades 5 and 7 for mathematics and reading comprehension. For vocabulary, however, the between-school differences in their average rates of growth were not significant. Prior ability and gender were not significantly related to growth rates for the multiwave models across all three subtests.

(3) The average bias of an unadjusted cross-sectional model against an adjusted cross-sectional model was substantial—about three months of schooling, whereas the average bias of an unadjusted multiwave model against an adjusted multiwave model was less than one-tenth of one month of schooling. The bias of two-wave models against multiwave models was also substantial.

(4) The estimates of school effects based on multiwave models were more reliable than those based on two-wave models. Although more reliable, they were still not very reliable (.76 for mathematics, .43 for reading and .14 for vocabulary).

Within-school sample size, reliability of the tests, and variability among schools are some of the factors that affect reliability.

The above findings lead to the following conclusions and practical implications.

(1) For the cross-sectional analysis, simple comparison of school means without controlling for students' intake characteristics is not appropriate for a fair and adequate evaluation of school effects. The bias of an unadjusted model against an adjusted model was quite large. In other words, if one does not control for differences among schools in their background characteristics, the estimates of school effects will be biased against less advantaged schools. This also suggests that it might be equally important to control the relevant school-level variables, even though this study was not able to include these school-level variables. Unless all the relevant variables at both student and school level are specified, the estimates would be biased especially for the cross-sectional models.

(2) The two-wave model did not estimate the students' growth rates reliably because of large measurement errors. The reliability estimates of adjusted differences between schools based on the two-wave model were lower than those based on the cross-sectional model. Nevertheless, this does not imply that the growth score is intrinsically unreliable. The reliability estimates based on multiwave models were much higher than those based on two-wave models. They were even higher than those based on cross-sectional models when there were

considerable differences among schools in their true rates of growth, as was the case for mathematics and reading. By reducing the measurement errors involved in the difference score, the multiwave model can estimate individual rates of growth more precisely and reliably. Since low reliability of the difference score has been one of the key problems of measuring change, it was interesting to see how much differences in reliability one additional data point can make. Moreover, precision of the estimates would be increased further if one had additional waves of data.

(3) Students' background variables were not significantly related to the rates of growth based on the multiwave model across all three outcomes. This suggested that one could use rates of growth as a performance indicator of school quality and get fairly accurate estimates of school effects without controlling students' background characteristics. This might be one of the conveniences of using multiwave models instead of using cross-sectional models. It would be practically better and easier for an evaluator to collect additional waves of data rather than to collect all the relevant background information that are related to selection into schools.

(4) The variance among schools in their effects was largest for mathematics and smallest for vocabulary. This means that schools vary most in their mathematics performance and least in their vocabulary performance. This is consistent with what one would expect given family influences on vocabulary. Since vocabulary is more related to students' home background rather than to specific school practices, the average rate of growth would not vary much among schools

after controlling for students' background variables. Because schools varied little in their adjusted growth rates in vocabulary, reliability estimates were low.

Because students' learning and growth is the primary goal of education and because growth can be reliably estimated for some subject areas, those estimating between-school differences should be advised to use rates of growth based on multiwave data as performance indicators instead of commonly used adjusted school differences based on cross-sectional models. This study examined the conceptual advantages of a multiwave model over cross-sectional or two-wave models. The empirical results were consistent with expectations. However, there are several points to be considered that were not addressed in this study.

First, this study did not include a wide range of students' background variables such as SES, or any school-level variables denoting specific school policies and practices. To make the comparisons as clear as possible, this study employed the simplest form of between-school differences. One of the interesting results of the study was that rate of growth based on multiwave model can estimate the school effects fairly accurately without controlling for student background. A study that includes more student-level variables, especially SES might be of interest. Moreover, this study estimated only Type A effects. It would be a useful extension to also estimate Type B effects. That is, a model that includes school-level variables describing school policy and context would be useful to help explain why schools vary.

Second, this study did not include the basic analysis of covariance model with

grade 6 scores used as covariates to adjust grade 7 scores. Because of the low reliability of the difference score between two data points, researchers have preferred the ANCOVA model to the two-wave model. The parameters for the ANCOVA model, however, could not be estimated with the HLM because of the multicollinearity between grade 6 scores and grade 7 scores. A possible way to solve this problem is centering the grade 6 scores around each school mean. But this study excluded this analysis because it was beyond the scope of this study. However, a comparison between the ANCOVA model and the multiwave model might be another interesting topic for future study.

Third, this study also employed the simplest form of multiwave model, that is, a three-wave model. The effects of schooling on students' rates of growth between grades 5 and 7 were only examined in this study. Therefore, it can only show the short-term effects of schooling on student outcomes. A future study that includes four waves or five waves of data can show different school effects, which are based on the long-term effects of schooling.

References

Aitkin, M. A., Anderson, D., & Hinde, J. (1981). Statistical modeling of data on teaching styles. *Journal of the Royal Statistical Society A, 144*, 419-461.

Aitkin, M. A., & Longford, N. T. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society A, 149*, 1-43.

Alexander, K. L. & Pallas, A. M. (1983). Private schools and public policy: New evidence on cognitive achievement in public and private schools. *Sociology of Education, 56*, 170-182.

Alexander, K.L., Cook, & McDill, (1978). Curriculum tracking and educational stratification. *American Sociological Review, 43(3)*, 47-66.

Barr, R. & Drebeen, R. (1983). *How schools work*. Chicago: University of Chicago Press.

Bereiter, C. (1963). Some persisting dillemmas in the measurement of change. In C. W. Harris (Ed.). *Problems in measuring change*. Madison, Wisconsin: University of Wisconsin Press.

Bridge, R. G., Judd, C. M., & Moock, P. R. (1979). *The determinants of educational outcomes*. Cambridge, MA: Ballinger.

Bryk, A. S. (1980). Analyzing data from premeasure/postmeasure designs. In S. Anderson, A. Auquier, W. W. Hauck, D. Oakes, W. Vandaaele, & H. I. Weisberg. *Statistical methods for comparative studies*. New York: Wiley.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear model to assessing change. *Psychological Bulletin, 101(1)*, 147-158.

Bryk, A.S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education, 97(1)*, 65-108.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. In D.C. Berliner (Ed.), *Review of Research in Education*, Washington, D.C.: American educational Research Association, 158-231.

Burstein, L., Miller, M. D., & Linn, R. L. (1979). *The use of within-group slopes as indices of group outcomes* (CSE Report Series). Los Angeles: Center for the

Study of Evaluation, University of California, Los Angeles.

Campbell, E. Q., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Coleman, J. S., Hoffer, T., & Kilgore, S. (1981). *Public and private schools. Report to the National Center for Educational Statistics.* Chicago: National Opinion Research Center.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Chicago: Rand McNally.

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

Dyer, H. S. (1970). Toward objective criteria of professional accountability in the schools of New York city. *Phi Delta Kappan, 52,* 206-211.

Dyer, H.S., Linn, R. L., & Patton, M. J. (1969). A comparison of four methods of obtaining discrepancy measure based on observed and predicted school system means on achievement tests. *American Educational Research Journal, 6,* 591-605.

Fennema, E. (1980). Sex-related differences in mathematics achievement: where and why. In L.H. Fox, L. Brody & D. Tobins (Eds.), *Women and the mathematical mystique.* Baltimore, MD: The Johns Hopkins University Press.

Glass, G.V., & Stanley, J. C. (1970). *Statistical methods in education and psychology.* Englewood Cliffs, NJ: Prentice-Hall.

Goldstein, H. (1987). *Multilevel models in educational and social research.* New York: Oxford University Press.

Gray, J. (1988). Multilevel models: issues and problems emerging from their recent appication in British studies of school effectiveness. In D. R. Bock (Ed.), *Multilevel Analyses of Educational Data,* New York: Academic Press, 1-19.

Haertel, E. H., James, T., & Levin, H.M. (1987). *Comparing Public and private schools, Vol 2: School achievement.* New York: The Falmer Press.

Haney, W. (1977). *A technical history of the national Follow Through Evaluation Vol. V. The Follow Through planned variation experiment.* Cambridge, Mass.: The Huron Institute.

Harris, C. W. (1963). (Ed.). *Problems in measuring change*. Madison: University of Wisconsin Press.

King, E. M. (1982). *Canadian Test of Basic Skills (teacher's Guide)*. Canada: Nelson Canada Limited.

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

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford, CA: The Stanford University Press.

Mandeville, G. K., & Anderson, L. W. (1987). The stability of school effectiveness indices across grade levels and subject areas. *Journal of Educational Measurement, 24(3)*, 203-216.

Marco, G. L. (1974). A comparison of selected school effectiveness measures based on longitudinal data. *Journal of Educational Measurement, 11(4)*, 225-234.

Martin, D. J., & Hoover, H. D. (1987). Sex differences in educational achievement: A longitudinal study. *Journal of Early Adolescence, 7(1)*, 65-83.

Murnane, R. J. (1987). Improving educational indicators and economic indicators. *Educational Evaluation and Policy Analysis, 9*, 101-116.

Raudenbush, S. W., Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59(1)*, 1-7.

Raudenbush, S. W., & Bryk, A. S. (1988). Methodological advances in analyzing the effects of schools and classrooms on student learning. In E. Z. Rothkopf (Ed.), *Review of Research in Education*, 423-475, Washington, D.C.: American Educationl Research Association.

Raudenbush, S.W. & Willms, J. D. (1988). *Sources of bias in the estimation of school effects*. Edinburgh University: Centre for Educational Sociology, and University of British Columbia: Centre for Policy Studies in Education.

Rogosa, D. R., Brandt, D. & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92*, 726-748.

Rogosa, D. R., & Willet, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement, 20*, 335-343.

Willet, J. B. (1988). Questions and answers in the measurement of change. In E. Z. Rothkopf (Ed.). *Review of Research in Education*, 345-422. Washington, D. C.: American Educational Research Association.

Willms, J. D. (1985). Catholic-school effects on academic achievement: New evidence from the High School and Beyond Follow-up study. *Sociology of Education, 58*, 98-114.

Willms, J. D., & Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement, 26*, 209-232.

Willms, J. D., & Jacobsen, S. (1990). Growth in mathematics skills during the intermediate years: sex differences and school effects. *International Journal of Educational Research*, 14, 157-174.

Wittrock, M.C. & Wiley, D.E. (Eds.) (1970). *The evaluation of instruction: Issues and problems*. Ney York: Holt, Rinehart, & Winston.

Appendix

Demographic Descriptions of the sample

	n	CCAT		CTBS math		CTBS read		CTBS vocab		Female
		mean	(SD)	mean	(SD)	mean	(SD)	mean	(SD)	mean
1	19	115.509	(10.678)	91.947	(7.982)	88.000	(8.888)	88.632	(10.377)	.474
				84.105	(5.537)	76.947	(8.229)	80.158	(8.604)	
				69.316	(9.019)	67.263	(9.291)	68.895	(8.906)	
2	38	109.382	(12.176)	82.576	(9.194)	80.342	(10.207)	82.105	(9.781)	.575
				74.632	(8.704)	71.289	(9.392)	75.632	(8.952)	
				60.868	(10.351)	60.632	(11.262)	62.316	(10.730)	
3	26	109.231	(13.955)	79.423	(13.949)	77.346	(13.532)	79.115	(9.425)	.538
				70.577	(11.024)	68.731	(12.824)	70.962	(8.417)	
				63.923	(10.677)	62.038	(10.698)	62.154	(9.922)	
4	39	108.739	(10.906)	83.026	(10.539)	79.359	(11.354)	80.103	(10.218)	.462
				74.590	(11.255)	72.641	(9.598)	73.385	(7.308)	
				62.667	(9.847)	59.846	(9.502)	63.000	(9.481)	
5	32	108.052	(14.618)	79.000	(12.297)	77.375	(12.045)	78.844	(11.274)	.531
				69.656	(9.973)	71..594	(10.320)	74.750	(9.253)	
				61.500	(11.769)	59.875	(9.366)	61.156	(10.287)	
6	42	107.960	(10.609)	82.786	(8,883)	80,095	(10.670)	81.238	(8.826)	.524
				73.857	(7.216)	70.071	(9.506)	74.048	(8.052)	
				61.905	(8.213)	62.548	(6.922)	63.738	(7.126)	

	n	CCAT		CTBS math		CTBS read		CTBS vocab		Female
		mean	(SD)	mean	(SD)	mean	(SD)	mean	(SD)	mean
7	26	107.679	(13.685)	82.692	(9.277)	80.346	(11.070)	79.731	(11.557)	.692
				72.115	(8.392)	71.077	(11.842)	72.423	(11.236)	
				63.846	(7.588)	62.038	(9.718)	60.296	(12.321)	
8	17	107.422	(11.299)	80.412	(13.496)	77.118	(8.184)	79.353	(10.099)	.647
				73.059	(9.430)	72.471	(8.676)	71.824	(7.519)	
				60.059	(7.636)	59.706	(9.218)	62.765	(8.671)	
9	38	106.833	(14.873)	79.237	(11.790)	78.526	(11.747)	76.342	(13.581)	.447
				71.842	(10.394)	71.184	(11.759)	70.842	(10.839)	
				61.974	(10.265)	58.158	(11.890)	60.263	(10.894)	
10	19	106.360	(13.210)	78.263	(10.572)	75.211	(10.628)	76.474	(11.172)	.526
				68.947	(10.926)	66.684	(10.149)	69.316	(11.036)	
				64.158	(9.002)	61.000	(10.451)	60.632	(10.383)	
11	24	105.910	(12.621)	84.458	(11.018)	78.375	(12.272)	75.833	(14.743)	.542
				77.167	(11.397)	71.250	(8.684)	71.833	(9.707)	
				62.917	(9.908)	59.583	(10.866)	60.625	(8.821)	
12	17	105.216	(12.510)	80.294	(10.658)	76.706	(10.209)	78.118	(9.911)	.647
				74.529	(10.248)	70.118	(10.222)	73.176	(11.293)	
				59.882	(8.746)	56.824	(10.858)	59.000	(8.746)	
13	16	104.979	(11.818)	80.062	(10.109)	78.562	(10.106)	79.937	(11.958)	.375
				72.312	(7.097)	69.312	(8.171)	69.625	(12.126)	
				57.625	(8.801)	57.437	(10.282)	58.375	(7.429)	

	n	CCAT		CTBS math		CTBS read		CTBS vocab		Female
		mean	(SD)	mean	(SD)	mean	(SD)	mean	(SD)	mean
14	26	104.917	(15.447)	74.346	(15.263)	73.269	(11.148)	79.192	(10.358)	.423
				67.038	(12.456)	66.538	(12.238)	71.038	(9.493)	
				59.615	(10.269)	58.885	(8.687)	59.000	(10.844)	
15	34	104.662	(16.167)	85.618	(9.474)	79.626	(13.660)	79.441	(11.368)	.588
				79.029	(8.997)	74.471	(7.684)	74.412	(9.494)	
				63.059	(10.245)	58.059	(13.071)	62.206	(12.715)	
16	16	103.615	(10.365)	81.812	(8.573)	78.437	(12.915)	77.937	(13.279)	.438
				71.687	(8.822)	68.250	(7.611)	72.875	(6.292)	
				61.437	(7.023)	58.062	(7.886)	60.812	(8.207)	
17	12	103.319	(14.344)	84.583	(10.613)	78.417	(9.020)	75.417	(10.630)	.500
				71.083	(10.122)	68.750	(11.522)	69.000	(11.176)	
				62.667	(10.057)	58.250	(8.476)	57.583	(12.280)	
18	22	102.977	(10.397)	80.909	(12.290)	77.409	(14.050)	78.409	(7.980)	.635
				69.545	(12.046)	69.000	(12.024)	71.545	(9.277)	
				59.864	(11.281)	57.273	(10.058)	63.273	(8.509)	
19	16	102.896	(15.933)	80.062	(9.740)	78.937	(10.096)	77.625	(10.059)	.250
				70.375	(10.589)	69.250	(9.462)	70.750	(10.618)	
				57.437	(11.314)	60.812	(12.117)	59.312	(9.707)	
20	55	102.730	(12.011)	82.400	(10.477)	78.545	(10.378)	78.800	(11.616)	.455
				76.709	(8.425)	70.764	(9.355)	74.036	(10.331)	
				60.836	(8.346)	58.673	(10.885)	60.509	(10.358)	

	n	CCAT		CTBS math		CTBS read		CTBS vocab		Female
		mean	(SD)	mean	(SD)	mean	(SD)	mean	(SD)	mean
21	36	102.509	(10.118)	85.000	(8.380)	80.083	(8.673)	78.750	(9.646)	.444
				72.917	(6.809)	68.778	(9.363)	70.278	(8.345)	
				60.777	(8.901)	57.389	(9.986)	59.222	(9.469)	
22	13	101.308	(11.987)	75.231	(7.247)	71.231	(10.353)	75.692	(7.642)	.538
				65.154	(10.953)	64.923	(8.311)	65.538	(7.891)	
				56.000	(7.937)	55.231	(10.895)	55.923	(10.851)	
23	12	99.500	(10.519)	75.833	(7.987)	72.667	(13.845)	72.833	(10.945)	.667
				65.417	(7.669)	63.750	(12.563)	67.583	(9.949)	
				53.917	(10.945)	55.250	(8.497)	54.583	(8.775)	
24	14	99.405	(10.580)	73.929	(8.704)	72.071	(6.650)	72.929	(10.759)	.500
				67.286	(6.696)	64.214	(6.5770	66.429	(6.734)	
				55.214	(4.388)	55.429	(5.854)	55.214	(5.563)	
25	19	97.009	(10.261)	75.316	(11.076)	69.316	(9.995)	73.947	(8.714)	.474
				67.316	(9.099)	64.947	(11.198)	68.316	(7.667)	
				54.158	(7.603)	51.579	(7.904)	56.632	(8.355)	
26	20	96.742	(11.830)	76.100	(9.684)	71.450	(10.324)	70.300	(8.921)	.450
				70.750	(8.595)	(12.792)	65.600	(8.535)		
				64.950	(7.783)	52.550	(13.221)	57.200	(7.811)	