RANKING BC SECONDARY SCHOOLS: A MULTILEVEL ANALYSIS APPROACH

by

STEPHANIE BARCLAY MCKEOWN

B.A., The University of British Columbia, 1996

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

MEASUREMENT, EVALUATION AND RESEARCH METHODOLOGY

THE UNIVERSITY OF BRITISH COLUMBIA

December 2004

Abstract

This study is an investigation of the current methods used to rank schools. An example used throughout this current study was the "Report Card on British Columbia's Secondary Schools", published annually by the Fraser Institute. However, the methodology applied for ranking schools concealled relationships within and between schools because it only included aggregate data. The interest for researchers lies not only in the average relationship between schools, but in how this relationship varies across schools. As an alternative approach, multilevel analysis simultaneously model disaggregate and aggregate data, which provides more information to the researcher about within and between school variance. The principal idea underlying the theoretical framework of multilevel analysis is that schools are hierarchical structures.

This present study adopted the multilevel assumption and aimed to investigate three research questions. First, how much of the variability in school performance on Grade 12 provincial examinations could be attributable to differences between schools and how much to differences within schools? Secondly, to what extent does the school attended influence the students' academic attainment? Thirdly, are there factors at the student and school levels that account for variability at either level? The findings in this study highlight how the sample of secondary schools in BC differed on examination achievement, and how including student-level information and school context allows researchers to identify the complicated relationships that occur within and between schools. The samples of schools were ranked according to their empirical Bayes estimates with 95% confidence intervals, which demonstrated that it was statistically invalid to compare a majority of schools based on the information collected in the present

study. The results from this study established a benefit of using multilevel models and the limitations to using report cards based on a single numerical score for comparing the differences between schools in BC.

Table of Contents

List of Tables

List of Figures

Acknowledgements

A special thank you and acknowledgement to Dr. Nand Kishor, my research supervisor and great educator, for his incredible patience and wonderful support throughout the time he has supervised my thesis. I also wish to extend my appreciation to Dr. Peter Gouzouasis and Dr. Valerie Overgaard for offering their valuable time, energy, and positive feedback. The research analysts at EduData Canada, Jennifer Lloyd and Andrea Hartshorne, are thanked for their help in providing me with the data, liaising with the Ministry of Education, and being so accessible whenever I needed their help. My sincere gratitude for the patience and support my loving husband, Jamie, has provided me over the years while working on my thesis project as well as taking such wonderful care of our 21-month old son, Dylan. I also wish to express warm recognition of my parents and sisters who have never lost faith in me, as well as my friends and colleagues, Deborah Matheson, Allan Hickey, Dawn Govan, Ken Burt, Lynda Lays, and Donna Blanleil, for their encouragement and emotional support.

For my family, my foundation…

"If you use catfish bait when you go fishing, you should not be surprised when you catch catfish. Those who are surprised are those who do not know catfish bait when they see it, or a catfish when they catch it." – Carver, 1975 (p. 86).

Chapter One

Statement of the Purpose and Problem

Introduction

Over the past four decades there has been much debate on the relevance and appropriateness of measures used to determine effective schools. Much of this discussion to date has taken place in Britain and the United States (US) (Aitkin & Longford, 1986; Austin, 1979; Edmonds, 1979; Goldstein, 1999; Goldstein & Rasbash, 1993; Gray & Jesson, 1990; Sammons & Nuttall, 1993; Scott & Walberg, 1979; Thomas, 2001). The debate began in 1966 when results from a research project commonly referred to as, "The Coleman Study," were released. The Coleman study identified schools as separate, self-enclosed and self-referential institutions. The findings led researchers to conclude that schools had little affect on students' achievement. Jencks and Brown (1975) also concluded that school characteristics contributed little to understanding the variation among student performance. Those were seminal studies, which involved the quantitative survey and administration of standardized tests to a large number of students and schools; they were influential in creating the foundation for a field of research activity called "school effectiveness research."

The school effectiveness research paradigm describes educational research concerned with exploring differences within and between schools. Researchers in this area focus on obtaining knowledge about relationships between explanatory and outcome factors using a variety of statistical models. What was troubling to early school effectiveness researchers who believed that there was an effect of schools upon student performance, was that Coleman et al. as well as Jencks and Brown's conclusions did not seem to be particularly

concerned with the impact of schools on the broad social systems in which they were embedded. Although school effectiveness researchers disagreed with their conclusions, they agreed that their findings were influential in setting the pathway for school effectiveness research (Goldstein & Woodhouse, 2000; Scott & Walberg, 1979). Carver (1975) suggested that Coleman et al.'s (1966) study motivated researchers to demonstrate that in prior work on school effectiveness, the effect of schools upon student performance had been neglected. In other words, the seminal research studies had concentrated on the effect of student performance on schools rather than schools upon student performance.

Recently, researchers investigating school effectiveness have attempted to demonstrate that, even when social and other factors were taken into account, there remained differences among schools which, they believed, could be attributed to the quality of the schooling process (Carver, 1975; Coe & Fitz-Gibbon, 1998; Cowley & Easton, 2004 Goldstein & Woodhouse, 2000; Raudenbush, 2004; Schmidt et al., 2001; Thomas, 2001). Edmonds (1979) believed that schools alone should be responsible for effectiveness and improvement. He argued that to emphasize personal characteristics in understanding student achievement would be a detriment to education, assigning too much responsibility to the family, and greatly diminishing the accountability of the school. Cowley and Easton (2004) agree with Edmonds. They argued that the more effectively a school enables all of its students to succeed, regardless of personal characteristics, the weaker the relationship will be between personal characteristics and academic success, thus placing the burden of responsibility back on the school rather than the family. Scott and Walberg (1979) agree with Edmonds in part, in that personal characteristics should be included when investigating what affects achievement. However, they added that the home, school, and

student all had significant roles to play in understanding school effectiveness.

The definition of school effectiveness has not been easy to operationalize, and thus measure. A number of different statistical techniques have been applied in school effectiveness research that have led to different conclusions about the importance of explanatory variables. When "The Coleman Study" was conducted in 1966, researchers believed that the construct of school effectiveness was correlated strongly with student achievement, and that student achievement could be measured by standardized assessments. Standardized assessments were viewed by the politicians as providing valid and interpretable comparisons among individuals and would avoid the introduction of personal bias by the measurer (Gibson & Asthana, 1998; Goldstein, 2001; Goldstein & Thomas, 1996). Coleman et al. (1966) theorized that two general areas - achievement and motivation - could define school effectiveness. In their study, they defined achievement as showing the accomplishments of the school to date, and motivation as showing the interest it has created for further achievements. They focused mainly on student achievement tests as the main outcome for measuring school effectiveness. The researchers argued that they could take average scores from students at each school, and could then compare these average scores across schools. However, they did not take into consideration any differences that may have already existed between schools.

The General Purpose

Recent British Columbia (BC) government policy implementing new systems of school accountability has highlighted the use of performance data to inform judgments about public schools. Furthermore, to stimulate school improvement there is the possibility that the allocation of public education funding, to some degree, could potentially be based on

'performance indicator' outcomes. Research on school effects has revealed important relationships of the school structure, school policies and teaching practices on school effectiveness (Gibson & Asthana, 1998; Goldstein, 1997; Goldstein et al., 1992; Gray et al., 1995; Nash, 2001; Scott & Walberg, 1979; Thomas, 2001; Yang & Goldstein, 1999; Waxman & Huang, 1997).

In Great Britain (GB), league tables ranking schools based on educational models are being produced (Goldstein & Spiegelhalter, 1996; Morrsion & Cowan, 1996; Thomas, 2001) and in the US, State Report Card tables are published to provide information on school effectiveness (Coe & Fitz-Gibbon, 1998; Lockwood et al., 2002; Zehr, 2001). Researchers applying multilevel modelling techniques to their analyses of school effectiveness research have stated, rather strongly, that league tables and rank-ordered tables are statistically invalid. They offer no practical understanding of how a school is effective and even less information on how a school has improved over time (Gibson & Asthana, 1998; Goldstein & Spiegelhalter, 1996; Goldstein & Thomas, 1996). These tables are used in GB and the US for high-stakes decision-making at the local authorities and district levels (Goldstein, 1997; Zehr, 2001).

In BC, as recently as March 2004, Cowley and Easton of the Fraser Institute produced a, "Report Card on British Columbia's Secondary Schools: 2004 Edition." They blended public and private schools into their analysis and compiled a rank-ordered table of the best to worst secondary schools across all regions in BC. The Fraser Institute publishes the report card on secondary schools: an independent economic, social and educational research organization. It is directed toward the parents and students as a tool to use in determining which school is the best for the student to attend rather than for government

accountability (Cowley & Easton, 2004). However, under the BC provincial government's newly implemented accountability framework, in which school districts enter into an accountability contract with the Ministry of Education, the possibility of school-to-school comparisons could be made based on the set of performance indicators defined in these contracts. Cowley and Easton (2004) suggested that school comparisons are at the heart of the improvement process, and that there is great benefit in identifying schools that are particularly effective. Nonetheless, research suggests that even where available performance measures are judged to be acceptable, there are inevitable limitations in making direct comparisons between schools (Gibson & Asthana, 1998; Goldstein, 1997; Goldstein et al., 1992; Gray et al., 1995; Thomas, 2001; Yang & Goldstein, 1999; Waxman & Huang, 1997, etc.). Moreover, the analysis underlying the BC report card does not use multilevel statistical models to examine the within-school and between-school differences, whereas multilevel modeling is now considered the most appropriate methodology for school effectiveness research (Yang & Goldstein, 1999; Nash, 2001; Schagen & Hutchison, 2003; Tekwe et al., 2004; Thomas, 2001). Hence, the findings from this present study will inform discussion on examining the inferences made about differences between schools reported in the 2004 Edition of the BC report card by using a statistical approach that incorporates school data as well as student level data.

The principal assumption underlying the theoretical framework of this current multilevel approach is that schools are not separate, self-enclosed, and self-referential institutions. Students are enrolled in schools, and schools are located in neighbourhoods, which house communities with changing characteristics. When communities change, neighbourhoods change, and schools are inevitably affected in some way. Therefore, on may consider that

we should not view students in seclusion, rather we should consider them in terms of having a nested relationship within a school, which is nested within a region, and all are nested in time. Information about individual schools and attributes about those schools are structured by time and location. In order to investigate these effects of nested students, schools, and regions, a statistical technique called multilevel modelling will be performed using student level and school level data. The fundamental importance of multilevel modelling is that substantive, interpretative benefits result from simultaneously modelling at several levels (Rasbash et al., 2002). The basic concept of this statistical modelling approach is that a variety of measures (i.e., scores from Grade 12 provincial assessments, school marks, etc.) at a lower level (i.e., student level) are nested within a higher level unit (i.e., the school). Multilevel analysis allows for an investigation of how school performance is influenced by the individual as well as the school (Goldstein, 1997). In consideration of these theoretical underpinnings, the purpose of this study is to take student characteristics as well as school contextual effects into account, and then based on the findings, discuss the validity of the inferences made from Fraser Institute secondary school rankings. If differences in student levels of achievement do differ across BC secondary schools due to school type, location, and characteristics, the potential effects of these differences will be discussed.

Research Questions

Cowley and Easton (2004) described school academic performance as a composite score out of 10, using eight indicators: (1) average provincial examination mark; (2) percentage of provincial examinations failed; (3) difference between school mark and examination mark in provincially examinable courses; (4) difference between male and female students

in the value of indicator #3, for English 12 only; (5) difference between male and female students in the value of indicator #3, for Mathematics 12 only; (6) provincially examinable courses taken per student; (7) graduate rate; and (8) composite drop-out rate.

The present study will address three research questions. First, it will investigate how much of the variability in school performance, on the overall provincial examination as well as on the English provincial examination, could be attributable to differences between schools and how much to differences between students within schools. Secondly, to what extent does the school attended influence the students' academic attainments? Thirdly, are there factors at the student and the school levels (e.g., gender, average years of parents' education, and school sector - public or independent) that account for the variability at either level?

Definition of Terms

The literature uses the term "school effectiveness research" to describe the investigation of within-school variation and between-school variations. However, there appears to be no theoretically coherent and generally accepted definition of school effectiveness. One reason why such a definition has not emerged is the tendency of researchers to consider many different statistical models to investigate schools and the types of performance indicators included in the analyses. There has been a propensity with school effectiveness researchers in the past to be narrowly focused on the task of ranking schools rather than on establishing factors, which could explain school differences. Also, there have been a number of researchers that have demonstrated that there are serious and inherent limitations to the usefulness of many performance indicators for providing reliable judgments about institutions and making comparisons across schools (Goldstein & Spiegelhalter, 1996;

Goldstein & Thomas, 1996; Goldstein & Woodhouse, 2000; Yang & Goldstein, 1999). School effectiveness research is an expanding field where many questions are left unanswered and still others are left unasked. The following are definitions of terms that will be used throughout this paper and are appropriate for the focus of this study.

School Effectiveness Research is concentrated on determining whether or not a school or school system is effective or ineffective. Schooling systems present an obvious example of a hierarchical structure, with students grouped within schools, which themselves may be clustered within education authorities (GB) or school districts (US and Canada). Educational researchers have been interested in comparing schools and other educational institutions, most often in terms of the achievements of the students. Such comparisons have several aims, including the aim of public accountability. However, in research terms, interest has recently been focused upon studying the factors that explain school differences (Goldstein et al., 1992; Opdenakker & Van Damme, 2000; Thomas, 2001; Sammons & Nuttall, 1993; Schagen, 1990).

Single-Level Models are the traditional linear methods such as those used by Coleman and his colleagues (1966), which measured relationships among student level variables, but ignored the actual ways in which students were allocated to schools and the influence of the school factors upon the students. These types of analyses result in two problems. The first is that the resulting statistical conclusions are often biased and overly optimistic. Traditional linear models offer a simple view of a complex situation that is statistically weak in its interpretation. This leads to the second concern that these models generally assume the same effects across groups, which fails to explicitly incorporate schools in the statistical model so that very little can be said about the influence of schools on student-

level variables (Goldstein, 1997). However, until fairly recently it was almost impossible to investigate the hierarchical nature of school data due to the lack of computer software applications sophisticated enough to handle multilevel models (Heck & Thomas, 2000; Schagen & Hutchison, 2003).

Multilevel Models are used on observational data with a hierarchical or clustered structure. Many of the populations of interest to educational researchers are organized into this type of structure. A hierarchy consists of units grouped and sometimes cross-classified at different levels. The analysis of nested data poses the unit of analysis problem, whereby deciding whether the analysis should focus on the individual or the group. Unfortunately, we often can't choose one over the other especially when looking at school effectiveness. Often the data are observations of individuals nested within groups. An assumption of group membership is that individuals within groups are more similar to one another than to individuals in other groups (Raudenbush, 2004). Multilevel models allow researchers to investigate different levels of analysis simultaneously, because they are concerned with the hierarchical structure rather than the individual level. From the variables specified at each level the program generates the linear model with the respective explanatory variables that account for response variability at each level. The hierarchical linear analysis not only estimates the model coefficients at each level, but also predicts the random effects associated with each sampling unit at each level. This can be empirically verified if the variance is partitioned so that the researcher can determine what proportion of variance is attributed to the individual and which proportion is attributed to the group (Heck & Thomas, 2000). It is hypothesized that if effects do differ across groups, differences can be explained with multilevel modeling.

Fixed Effects are variable coefficients that are constant across groups and do not vary. It is assumed that in the present, study gender exerts the same (fixed) impact within each school and would have a similar affect on student performance across our sample of schools. Therefore, gender would have a fixed effect on student performance. ESL status is also fixed in this study.

Random Effects are variable coefficients that can vary across groups. In other words, the coefficient can take a different value for each group (i.e., school). In the current study it is assumed that students' school marks could impact performance as a random variable: as such its impact on performance will vary across the schools in the sample.

Justification for the Study

The recent members of the school effectiveness research community have devoted considerable effort to understanding the variables that influence school effectiveness (Nash, 2001; Tate, 2004; Thomas, 2001; Yang & Goldstein, 1999). In the climate of accountability so prevalent in education in this new millennium, patterns of evidence must be provided to substantiate summary or evaluative conclusions. The multilevel analysis approach offers a method to uncover underlying trends and levels of effects within the hierarchical organization of schools. Multilevel analyses provide important empirical descriptions that are sensitive to contextual and environmental effects varying by location and time. The two-level multilevel model to be developed in this study also allows one to gauge the relative importance of background and outcome variables to the understanding of school effectiveness.

Finally, as mentioned earlier, considerable methodological difficulties have plagued research on school effectiveness (Goldstein, 1997; Goldstein & Woodhouse, 2000;

Raudenbush, 2004; Scheerens et al., 2001; Townsend, 2001; Willmott, 1999). Traditional educational single-level statistics failed to provide appropriate techniques to estimate the effects of performance scores on school effectiveness (Goldstein, 2001). Goldstein and his colleagues have been working on many research projects in Great Britain, especially with the schools in the Inner London Education Authority. There has been little effort in incorporating a systemic statistical model for describing school effectiveness by traditional researchers in the field. The methodological significance of this study is that it demonstrates the power of multilevel analysis to educational research and provides a much stronger theoretical framework for reviewing how schools differ compared to traditional linear methods. Multilevel models can be used to estimate the effects of particular background variables that seem to be significant regarding the effectiveness of a school.

Chapter Two

Review of the Literature

The purpose of this literature review is to form a context in which school effectiveness research can be understood theoretically and practically. This review is divided into four sections. The first provides the historical development and criticisms of school effectiveness research. The second section briefly provides a discussion on measurement, and why some contextual and environmental effects have been overlooked or purposively ignored by the pioneers of this field of research. The third section describes the level of analysis dilemma and identifies multilevel frameworks for investigating school effectiveness as inherently hierarchical or multilevel structures. The chapter concludes with a discussion on performance indicators and school comparisons as they pertain to the research questions in this paper.

The Historical Development and Criticisms of School Effectiveness Research

As was briefly described in Chapter One, the initial debate on how to measure school effectiveness began in 1966 when a seminal research project was conducted on behalf of the National Center for Educational Statistics of the US Office of Education (Austin, 1979; Coleman et al., 1966; Gibson & Asthana, 1998). James Coleman and his colleagues were involved in the design, administration, and analysis of the study (Coleman et al., 1966). The study was based on a nation-wide survey, "Equality of Educational Opportunity," that was administered to elementary and secondary school students and principals. Standardized assessments were administered to all students to test academic achievement. Using regression analyses of scores from the standardized assessments, Coleman and his colleagues concluded that of the total variance explained, they attributed only 10% to

between school differences, compared with 90% attributed to individual differences within schools. At the time of the study, class and ethnicity segregated the majority of schools in the US.

The findings from those regression analyses led Coleman and his colleagues to conclude that the social class/prior achievement mix of schools was the only school variable that had any impact on academic outcomes. It was the main variable they found to be particularly influential (Coleman et al., 1966). The family – not the school – was the major variable in determining achievement. They also concluded that of all the school factors that had the greatest influence, the teacher's characteristics were more significant than the school facilities or curriculum (Austin, 1979). Almost a decade later, Jencks and Brown (1975) also argued that schools did not play a large role in understanding the variance in academic performance. Using data from Project Talent, a longitudinal study of students in grades nine through twelve in 1960, they suggested that changes in high school characteristics such as teacher experience, class size, and social composition were unlikely to change high-school effectiveness. They argued that the data supported a focus on differences within high schools rather than differences between high schools (Jencks & Brown, 1975).

The pioneers of school effectiveness research did not propose that schools had no effect. Rather, they indicated that when investigating differences in the effect of schooling between schools, it was difficult to identify school-related variables that accounted for the observed differences (Austin, 1979). Gibson and Asthana (1998) suggested that from these major early findings, the US government shifted its political focus toward the social mix of schools. In the US this led to 'bussing', the physical transfer of students between schools in an attempt to create more socially and racially balanced school populations. A similar

intervention took place in GB in order to create more socially mixed class populations in schools (Gibson & Asthana, 1998). In both GB and the US, this political movement proved to be very problematic and other school effectiveness researchers began to question whether they could interpret the conclusions formed by Coleman et al. as the only intervention available to educational practitioners (Carver, 1975; Gibson & Asthana, 1998). In contrast to the seminal studies, Edmonds held the schools alone responsible for improvement (Edmonds, 1979; Scott & Walberg, 1979). He asserted that by emphasizing the background characteristics of students, the educators would have limited responsibility to be instructively effective, and it would put a heavy burden on parents. With similar reasoning, Cowley and Easton (2004) concluded that schools do matter. They argued that variations in student performance between schools should not be attributed to personal or family characteristics. In other words, family characteristics may be related to student academic performance, but that an effective school will enable all students to succeed, thus reducing the strength of this relationship significantly (Cowley & Easton, 2004). Schmidt et al. (2001) also argued strongly that schools matter, based on data on curriculum and achievement from a cross-national study – the Third International Mathematics and Science Study (TIMSS). In the middle of the continuum of whether schools matter or not, Scott and Walberg (1979) theorized that the student, the school, and the home were like a three-legged stool that was as strong as its weakest leg. They reasoned that strengthening the stronger legs was far less productive than strengthening the weakest. Therefore, they would have argued by including multiple indicators to describe student, school, and home characteristics, the data analyses would have more discriminative power to determine what contributed to school effectiveness.

In the late 1970s and 1980s, both the US and GB were looking to the efficiency of schools, their value for money, and their effectiveness in achieving measurable goals (Gibson & Asthana, 1998; Goldstein, 2001; Willmott, 1999). In GB, the view of the government was to provide enough information to parents so they could compare schools when choosing which ones their children would attend. Policy makers believed that schools could be compared if a standard measure, similar to a yardstick to measure distance, were to be applied to all schools equally (Goldstein, 2001). Proponents of standardized assessments considered them to be the most objective measures because standardized assessments were more impartial than teacher-constructed tests due to their distance from instruction. Also, they argued that a standardized measurement could be accepted as providing valid and interpretable comparisons among individuals, as they could avoid the introduction of personal characteristics of the measurer into the process (i.e., the individual teacher) (Goldstein & Thomas, 1996).

Rising criticisms of only using standardized assessment outcomes emerged exposing the limitations of the data and the findings from those types of studies. Carver (1975) argued that using standardized assessments that were inappropriately designed to measure school effectiveness, like the ones used in "The Coleman Study", would result in meaningless findings. Further criticisms suggested that studies that used only one year's outcomes measures were unable to adjust for intake differences and thus were rather uninformative (Aitkin & Longford, 1986; Carver, 1975; Coe & Fitz-Gibbon, 1998; Goldstein & Thomas, 1996). This latter argument was the basis for the development of an approach known as value added outcomes (Aitkin & Longford, 1986; Gibson & Asthana, 1998; Goldstein et al., 1992; Goldstein et al., 1996; Gray et al., 1995). Value-added outcomes are measures of

academic performances that provide baseline information to the researcher in which to measure the value added by schools in the learning process of students.

There was a growing awareness among school effectiveness researchers that in order to investigate school effectiveness, it was necessary to determine how much value the school added to the learning process (Gibson & Asthana, 1998; Schagen & Hutchison, 2003; Thomas, 2001). This required measures of prior achievement against which to compare educational outcomes in terms of gains or growth (Gray et al., 1995; Opdenakker & Van Damme, 2000). Raudenbush and Willms (1995) identified two types of effects to distinguish value added results. Type A effects controlled only for student level factors, including prior attainment and limited background variables, while Type B effects controlled for school level and contextual factors outside the school's control (Schaegen & Hutchison, 2003). Type A comparisons generally make little or no attempt to understand the effect of the schooling process, whereas Type B comparisons would be more useful to researchers interested in the effects of school policy on achievement. The Type A comparisons for school effectiveness studies that adjust only on achievement at entry to the school, may be inadequate. A common feature of value added analysis is differential effectiveness. Schools may achieve quite different results for the initially low achievers as compared with the initially high achievers (Hopkins et al., 1999; Opdenakker & Van Damme, 2000; Opdenakker & Van Damme, 2001; Pituch, 1999; Sammons & Nuttall, 1993). There may also be differential effectiveness across regions (Thomas, 2001; Webster & Fisher, 2000). Goldstein and Spiegelhalter (1996) drew the distinction between the value added model and adjusted comparison model. They argued that justifying the number of units an institution has added to its students is impractical, because depending

on the nature of the study, what determines value is relative. However, given that the variables included in the investigation have similar prior baseline measurements and measure the same contextual variables, these would be referred to as an adjusted comparative model. In order to incorporate the value added or adjusted model into educational research, more and improved data was required at the individual level. As a result, there was a heightened demand for student level data, both on intake and outcome, in determining school performance (Aitkin & Longford, 1986; Gibson & Asthana, 1998; Goldstein, 2001; Schagen & Hutchison, 2003; Wyatt, 1996).

Over the years, there have been improvement and significant contributions made to the school effectiveness paradigm. In the 1980s, a special type of statistical model called multilevel (which was developed in GB, and somewhat simultaneously, hierarchical analysis was developed in the US) greatly improved the way in which effective schools analyses could be handled and investigated (Aitken & Longford, 1986; Ballou et al., 2004; Bickel & Howley, 2000; Goldstein, 2001; Gray et al., 1995; Hopkins et al., 1999; Kezar & Eckel, 2002; Mok, 1995; Nash, 2001; Paterson & Goldstein, 1991; Pituch, 1999; Wong, 1996). Some of these improvements can be attributed directly to the many strong opponents to this type of research activity (Coe & Fitz-Gibbon, 1998; Slee & Weiner, 2001; Thrupp, 2001a; Thrupp, 2001b; Townsend, 2001b; Walford, 2002; Willmott, 1999; Wyatt, 1996). Through reviewing the literature, three overarching criticisms were revealed: (1) school effectiveness research has not been rooted in theoretical rationale with a common understanding and clear definition of effectiveness, or how it should be measured (Coe & Fitz-Gibbon, 1998; Hargreaves, 2001; Thrupp, 2001a); (2) school effectiveness research has been oversimplified by neglecting to include appropriate

contextual variables in the model (Coe & Fitz-Gibbon, 1998; Thrupp, 2001a); and (3) school effectiveness research has been unable to control the political use of its findings (Slee & Weiner, 2001; Thrupp, 2001a; Willmott, 1999; Wyatt, 1996). In response to these criticisms, school effectiveness researchers accepted, for the most part, that if school effectiveness research was going to mature, there needed to be a conceptual shift in the framework for this field of research (Freeman et al., 1998; Goldstein, 1997; Hargreaves, 2001; Walford, 2002; Wyatt, 1996).

Under-theorizing of School Effectiveness Research. Coe and Fitz-Gibbon (1998) addressed criticisms raised by leading exponents of qualitative research who challenged the school effectiveness paradigm stating it is a mechanistic and instrumentalist view of the schooling process. Some have argued quite strongly that school effectiveness research continually produced poor quality research (Gibson & Asthana, 1998; Goldstein & Woodhouse, 2000), and many have noted that it has not provided solutions for school improvement (Freeman et al., 1998; Hargreaves, 2001; Hopkins, 1999). The main argument is that the validity of interpreting school effectiveness as the true effectiveness of a school depends, in large part, on the choice of indicators measured (Cistone & Bashford, 2002; Cronbach & Linn, 1997; Gibson & Asthana, 1998; Lane & Stone, 2002; Linn, 2001). Researchers have emphasized that the choice of outcomes was often data-driven, motivated by availability of data rather than the desire to measure what was important (Coe & Fitz-Giboon, 1998; Goldstein & Woodhouse, 2000).

In response to the criticism that school effectiveness research was not encased in theory, Goldstein and Woodhouse (2000) suggested that instead of fitting the theory to school effectiveness research, a greater concentration on the definition of school effectiveness was

more prominently required. They recommended changing the term school effectiveness to educational effectiveness, and changing school improvement to institutional change. Coe & Fitz-Gibbon (1998) also found the term school effectiveness to be misleading and recommended that it would be more appropriate to talk of the adjusted academic performance of specific groups. Perhaps by changing the names, they would argue that the focus was no longer on the school but more so on the process of schooling, which would involve all aspects of education, not just the cognitive aspects (Wyatt, 1996).

Raudenbush and Willms (1995) proposed a theoretical model that included processes. They declared that student performance outcomes were influenced by three general factors: 1) student background characteristics, 2) school context (Opdenakker & Van Damme 2000; Opdenakker & Van Damme, 2001), and 3) processes of school policy and practice (Pituch, 1999; Waxman et al., 1997). Along a similar line of reasoning, Opdenakker & Van Damme (2001) argued that school composition (i.e., mix) should be included in the second factor, school context. In their study, they demonstrated that by neglecting to include the relationship school composition had with mathematics achievement, the effect of school processes was overestimated.

Raudenbush and Willms' (1995) three-factored conceptual framework is comparable to Scott and Walberg's three-legged stool model. The main advantage of this theoretical approach was that it allowed researchers to identify institutions whose practices appeared to promote evidential outcomes (academic outcomes), while controlling for contextual variables (i.e., the social and economic conditions within which students and schools function) that generally lay outside the control of administration and faculty (Goldstein, 1997). Similarly, Hargreaves (2001) offered an in-depth discussion of a capital theory of

school effectiveness in which he described four master concepts and two sub-concepts for each: intellectual capital, which includes the creation and transfer of knowledge; social capital, generating trust and maintaining networks; educational outcomes, both cognitive and moral; and leverage strategies that are grounded in evidence-informed and innovative practice. Wyatt (1996) also argued that the investigation of effectiveness needed to be broadened to include social outcomes from schools, which may be independent from academic outcomes.

Oversimplification of School Effectiveness Research. The theory of school effectiveness has also been criticized for over-claiming the role of the educational institution in student performance (Coe & Fitz-Gibbon, 1998; Scott & Walberg, 1979; Thrupp, 2001). It has been charged with excluding particular contextual variables that could significantly bias the relationship between student and school (Coe & Fitz-Gibbon, 1998), and that researchers have downplayed the relationship between social class and student achievement because they have been more interested in the school as a vehicle rather than the students within it.

Thrupp (2001a) and Wyatt (1996) claim to have reviewed many school effectiveness studies and found that the effect size of schools on student learning was quite small. However, they may have been limited in the types of research studies they included in their analyses. In a review of the field, Teddlie and Reynolds (2001) found that there were three distinct strands of school effectiveness research: (1) school effects research, (2) effective schools research, and (3) school improvement research. Researchers interested in school effects only scientifically study the relationship between the school and student performance (Aitkin & Longford, 1986; Austin, 1979; Jencks et al., 1975; Kezar & Eckel, 2002; Raudenbush & Willms, 1995; Schagen, 1990; Scott & Walberg, 1979). Effective

schools research incorporates school processes and contextual effects, such as classroom, teacher, environment and time (McCaffrey et al., 2004; Opdenakker & Van Damme, 2000; Opdenakker & Van Damme, 2001; Paterson & Goldstein, 1991; Sammons & Nuttall, 1993; Thomas, 2001; Waxman et al., 1997). School improvement researchers would state that rather than attempting to determine what is an effective school, it is more relevant to determine what causes particular schools to improve in relation to other schools (Freeman et al., 1998; Goldstein et al., 1992; Goldstein, 1997; Gray et al., 1995). Goldstein (1997) and Gray et al. (1995) would caution that researchers might be able to determine whether differences existed between schools, but could not determine how well a particular school was performing with any precision.

The further development of effective schools research has exposed cross-classified data structures. These are found when units of analysis are not purely one level within the hierarchical structures. Raudenbush and Bryk (2002) provided an example of children attending a set of schools and living in a set of neighbourhoods. This is not a simple hierarchical structure where all the students in a particular classroom attend only one school, but rather where the school draws from multiple neighbourhoods and where children living in a neighbourhood attend multiple schools. It appears then, that the oversimplification criticism is already being addressed in the research.

Political Relationship. The continued interest in investigating effective schools is obvious, especially in light of stakeholder and policy makers' desires to hold schools accountable for their role in increasing student and learning performance. However, another major criticism of the school effectiveness paradigm is that it has not controlled its findings for political uses, and that it has developed an intimate relationship with politics across school,

local, and national levels (Coe & Fitz-Gibbon, 1998; Goldstein, 2001; Willmott, 1999; Wyatt, 1996). Slee and Weiner (2001) suggested that school effectiveness research has distanced itself from controversies about educational policies that shape school processes and outcomes. They argued that the term "value added" is saturated in political overtones, and that the school effectiveness researchers have openly ignored it. This criticism has potency in that politicians and educational stakeholders bought into the theory of institutional effectiveness, but tended to use the results inappropriately for school-to-school comparisons (Goldstein, 1997; Goldstein, 2001; Goldstein & Speigelhalter, 1996; Slee & Weiner, 2001; Thrupp, 2001). Recently, school-to-school comparisons have become associated with achievement targets and ultimately as a means of allocating resources towards individual schools (Chester, 2003; Goldstein, 2001; Schafer, 2003; Zehr, 2001). Zehr (2001) reported that the Indiana State Board of Education has adopted a plan that ranks schools based on their students' scores on state assessment that will be implemented for the 2005 academic year. They intend to reward high performing schools, and will provide technical aid and resources to low performing schools. Despite this, Thrupp (2001) and Slee and Weiner (2001) were too inclusive by indicating that all school effectiveness research is politically motivated. Many school effectiveness researchers would argue that when policy makers develop policy and offer rewards based on changes in standards that are not supported by empirical data, they have no way of knowing whether what is observed is real or not because their assumptions could be wrong (Aitkin & Longford, 1986; Goldstein, 2001; Goldstein & Speigelhalter, 1996; Hopkins et al., 1999; Morrison & Cowan, 1996; Reynolds & Teddlie, 2001; Scheerens, et al., 2001).

Supporters of school effectiveness research advocate that it is a difficult and

complicated field of work where someone could always find something to criticize (Goldstein, 2001; Scheerens et al., 2001). Coe & Fitz-Gibbon (1998) appropriately wrote, "however, the stakes are high, particularly for schools in current political climates, so a clear acknowledgement and summary of these problems is important (p. 11)." Goldstein (1997) invited proponents of school effectiveness research to welcome criticisms and learn from them so that the pedagogical discourse on school effectiveness and improvement can develop and mature. In spite of Thrupps' (2001a) complaints that school effectiveness researchers have not responded to critics, many have attempted to respond and have taken the opportunity to contribute to the intellectual and practical development of the concept and theory. The main assertions extracted from the early development of this field of research and the comments and recommendations from critics suggest that further work must be conducted in determining: a theoretical framework, operational definition, and appropriate methods for investigating school effectiveness and improvement research; how to obtain reliable and valid data for analyses, rather than what is accessible; and the most appropriate statistical procedures to use when analysing the data.

Measurement Concerns

Goldstein (1997) acknowledged measurement issues in his paper, but only briefly. It is important to include a discussion on educational measurement and theory because school effectiveness researchers have been criticized for focusing too much on issues about measurement, while neglecting to develop a theory (Slee & Weiner, 2001; Thrupp, 2001b). Most researchers would agree that measurement does not occur in a theoretical vacuum. Therefore, meaning must first be provided and then an explanation should be derived that is contained in theoretical qualitative or quantitative measures. In his development of a new

capital theory for school effectiveness, Hargreaves (2001) described a useful theory as one that "contains a relatively small set of concepts in explicit relationships, and measured variables should be capable of being contained within the concepts (p. 487)." This implies that measurement of individual variance assumes individuals belong to a definable population and can be assigned a number to an attribute according to a rule, as to infer the location of the individual in the attribute distribution of the entire population (Bock, 1989; Michell, 1999). In other words, educational measurement is often concerned with studying variables or constructs that are not observed directly, but can be measured by a range of indicator variables assumed to be associated with the construct (Morrison & Cowan, 1996). Consider the construct socio-economic status, which cannot be directly observed. However, it could be measured from information obtained on four indicator variables: father's education, mother's education, father's occupation, and family income (Cowley & Easton, 2004; Morrison & Cowan, 1996). Keeping in mind, that there must be proven validity of the types of indicators measured that are demonstrated, not just assumed. Also, consideration at the various levels (classroom and school) at which changes in performance can be explained by relevant variables (changes in funding, enrollment, instruction, etc.) will help strengthen the evaluation of the impact of performance indicators (Burke et al., 2002; Lane & Stone, 2002; Goldstein, 2001; Goldstein & Thomas, 1996).

When studying what the validity of inferences made from responses and performances means in psychological and educational research, one of the best sources of information on the topic is Samuel Messick. He argued strongly that the inferences made from the results of performance assessments need to be systematically addressed because of the potential consequences that could occur based on these inferences (Messick, 1995). Messick

identified two sources of invalidity: construct under-representation, and construct-irrelevant variance. The first source refers to inferences made from an assessment that was too narrow and failed to include important relationships of the construct. The second source refers to inferences made from an assessment that was too broad and included information that was indirectly related to the construct through other irrelevant variables. The first source of error is of particular interest for the current study, in that rank-ordered school tables typically fall into this category by incorporating only one or relatively few indicators of school performance.

Educational researchers have been interested in comparing schools most often in terms of outcomes – the achievements of their students. Such comparisons had several aims, including the aim of public accountability (Goldstein, 2001) but, in research terms, interest was usually focused upon studying the factors that explained school differences (Goldstein et al., 1997). The underlying goal has been to isolate those characteristics, which distinguish effective schools from the rest (Austin, 1979; Bickel & Howley, 2000; Cowley & Easton, 2004; Gray & Jesson, 1990; Opdenakker & Van Damme, 2001; Sammons & Nuttall, 1993). The fundamental rationale is that the responsibility for school performance, and thus for school improvement, rested with individual schools – their staff, administrators, board members, parents and students. Rather than determining the ingredients for an effective school, Reynolds and Teddlie (2001) and Freeman et al. (1998) recommended that researchers focus on describing why a particular school is effective, which would include an investigation of contextual factors (Gibson & Asthana, 1998; Gray et al., 1995; Nash, 2001; Raudenbush & Willms, 1995; Schmidt et al., 2001 Yang & Goldstein, 1999; Waxman & Huang, 1997). However, as mentioned earlier, Goldstein

(1997) and Gray et al. (1995) cautioned against this type of research because they argued that determining why a school is effective could not be decided with any precision. Thus, from a research perspective, the degree of uncertainty that would be tolerated would depend on the nature and purpose of the study (Cronbach & Linn, 1997; Messick, 1995; Michell, 1999). A methodological dilemma that has plagued school effectiveness research has been the level or unit of analysis to select in the study (Goldstein, 1997; Goldstein & Woodhouse, 2000; Raudenbush, 2004; Scheerens et al., 2001; Townsend, 2001; Willmott, 1999).

Level of Analysis Dilemma

As identified in the previous sections, a number of different modeling procedures have been used under the school effectiveness framework that have led to different conclusions about the importance of individual explanatory variables and the estimated effectiveness of institutions. These differences have led to arguments over the appropriate level or unit of analysis to investigate under these models (Goldstein, 1997; Goldstein & Thomas, 1996; Heck & Thomas, 2000; Raudenbush & Bryk, 2002). To put it in perspective, consider the following question: when the object of analysis is the assessment of the importance of school level variables, but we have student level outcomes, should these be aggregated to the school level for the analysis, or should we analyse the individual outcomes? If the latter is the answer, Aitkin and Longford (1986) would then ask, how should the school structure be represented in the model? If the former is the answer, then how much information would the researchers be willing to lose? This multilevel structure makes it very difficult to determine the appropriate level or unit of analysis. Differences in unit selection will lead to different conclusions, sometimes contradictory, on understanding and

explaining the variances between school performances (Paterson & Goldstein, 1991; Raudenbush & Willms, 1995; Sammons & Nuttall, 1993; Schagen & Hutchison, 2003).

Single Level Models. Traditional linear methods, such as those used by early school effectiveness researchers (e.g., Coleman et al., 1966; Jencks, 1975) were interested in relationships among student level variables, but neglected to investigate how students were allocated to schools (Goldstein, 1997; Raudenbush & Bryk, 2002). Schools represent an obvious example of hierarchical data with students grouped into classes that are clustered within schools. Past research strategies for dealing with these complex multilevel structures has been limited. This is partly due to the unavailability of the estimation procedures used to analyse these datasets (Schagen & Hutchison, 2003; Raudenbush & Bryk, 2002), as well as the possible lack of consideration for the implications of assumptions made about measuring variables within their natural level, or about moving them to another level through aggregation or disaggregation (Heck & Thomas, 2000). Aggregated datasets include individual level data that were aggregated to the higher level for analysis (i.e., school level). A simple illustration of unit selection can be explained within the context of a study that is investigating what affects student achievement in a sample size of 50,000 students attending 65 schools. If we used aggregated student-level data, we would end up estimating the standard errors for the school level using the aggregated 50,000 student cases, but would end up with only 65 aggregated observations. The information available in student level data would be lost or diminished in aggregate level analyses, because the variability in performance of each individual would be reduced to a single school level variable (Heck & Thomas, 2000). Failing to acknowledge the within group variability in the dataset can possibly distort the relationships examined

between units by either overestimating or underestimating the effects (Aitkin & Longford, 1986; Bickel & Howley, 2000; Gibson & Asthana, 1998). Schagen & Hutchison (2003) suggested that the biggest impact that educational research has had on education and society over the past 20 years has been the combination of a sophisticated analysis technique and the availability of databases in which to apply the techniques for analyzing multilevel data structures.

Multi Level Models. People tend to exist within organizational structures, such as families, schools, and business organizations. These structures are multilevel phenomena in which individuals are nested within clusters such as classrooms, departments, neighbourhoods that are also nested within larger units such as schools, businesses, cities, which are all nested in time and location. Each nested unit may also interact with contextual factors and characteristics of the lower level or upper level of the structure. For example, student learning generally takes place in the classroom, which could be affected by the quality of instruction (Ballou et al., 2004; Blatchford et al., 2002; Goldstein, 2001; McCaffrey et al., 2004), curriculum (Schmidt et al., 2001), class size and processes (Blatchford et al., 2002; Waxman & Huang, 1997), school climate or culture (Kezar & Eckel, 2002; Opdenakker & Van Damme, 2001; Pituch, 1999; Schagen, 1990; Thomas, 2001) and other contextual variables. Most early school effectiveness initiatives poorly conceptualized the many ways in which these contextual variables could impact upon the learning and performance at the student level, school level, or school district level (Coe & Fitz-Gibbon, 1998; Thomas, 2001). Within a multilevel framework, indicators selected should provide a detailed description of the school without neglecting the hierarchical context structured in location and time, as well as keeping in mind how the indicators collected appropriately express the

internal and external values of effectiveness for meeting provincial, district, and institutional needs (Burke et al., 2002).

Consider the example in British Columbia (BC), where provincial examination results for students enrolled in grade 12 academic programs at the end of the schooling year are collected for each school in the province. In BC, students generally write their grade 12 provincial examinations in June of each year, but are also able to write or re-write in November and January. A school effectiveness researcher would want to know whether a particular kind of teaching practice in some schools is associated with improved examination performance for certain subjects. The researcher also would hopefully have good measures of the students' achievements and some background characteristics when they started the period of schooling so that she or he could control for in the analysis. The traditional approach to the analysis of these data would be to carry out a regression analysis, using performance scores as response, to study the relationship with teaching practice, adjusting for the initial abilities (intake measures). A straightforward regression analysis lacks process validity by failing to take account of the school level variables and the clustering of students in the classroom. These models generally assumed the same effects across groups, which failed to explicitly incorporate schools in the statistical model so that very little can be said about the influence of schools on student level variables (Goldstein, 1997). If standard errors are underestimated, it might be inferred that findings are statistically significant when they are not. Traditional linear models offered a simple view of a complex situation, which was statistically weak in its interpretation because so much information was lost or diminished. However, a multilevel analysis accounts for these clusters. Individual students grouped together in a school or classroom share

common experiences, which make their results more homogeneous than those of a random sample of students drawn from the population of all schools. This greater homogeneity is naturally modeled by a positive within-school correlation among student results in the same school (Aitkin & Longford, 1986). An analysis that explicitly models the manner in which students are grouped within schools has several advantages. First, it enables the researcher to obtain improved estimation of individual effects at each level of analysis. Secondly, by modeling cross-level effects (how variables measured at one level affect relations occurring at another) using the clustering of information it provides correct standard errors, confidence intervals and significance tests, which will generally be more conservative than the traditional regression analysis. Thirdly, multilevel analysis allows researchers to vary intercepts and slopes. The model partitions the variance, which is attributed to the individual and to the group. By partitioning variance-covariance components (partitions learning rate variance into within and between school components) it allows the use of covariates to be measured at any of the levels of hierarchy. The primary importance of multilevel modeling is that substantive, interpretative benefits result from simultaneously modeling at several levels. Multilevel analysis allows for an investigation of how student performance is influenced at the individual level as well as the school level (Aitkin & Longford, 1986; Goldstein, 1997; Osborne, 2000).

Performance Indicators and School Comparisons

Performance reporting has become the preferred approach of provincial and state policy makers on accountability for public higher education (Burke et al., 2002; Chester, 2003; Cordeiro & Vaidya, 2002; Gray, Goldstein & Thomas, 2001; Linn, 2001; Schafer, 2003). Accountability is used here as referring to the demonstration of a school's performance

measured in quantifiable terms to a public or political audience (Cistone & Bashford, 2002). Politicians rationalize these types of performance reports as demonstrating accountability, providing information for improving school performance, and for meeting provincial or state policy needs. In light of current political conversation regarding institutional accountability, there is the possibility that the allocation of public funding, to some degree, could potentially be based on the reporting of performance indicator outcomes (Lockwood et al., 2002; Zehr, 2001). Performance indicators can be defined as a summary of statistical measurements on an institution or system which are intended to be related to the 'quality' of its purpose (Goldstein & Spielgelhalter, 1996). In the education sector, such measures concern different aspects of the school or educational system. There are 'intake' indicators that include prior performance and student background characteristics; 'input' indicators such as the student/teacher ratio used to estimate the resources available to schools, 'process' measures such as the number of student enrollments to reflect organizational structure, and 'outcome' measures such as credentials awarded or student performance on assessments to success, completion and graduation rates (Burke et al., 2002; Goldstein & Spielgelhalter, 1996). Morrison and Cowan (1996) argued that there was a distinction between reporting the results from performance indicators in performance tables compared with rank-ordered tables. The former would more appropriately be defined as a profile that includes quantitative as well as qualitative information to describe the schools. However, there has been a propensity with some school effectiveness researchers and politicians to be narrowly focused on the task of ranking schools rather than on establishing school profiles (Morrison & Cowan, 1996). Some have used performance data, unadjusted for intake or context, for comparing

achievement across schools and reporting these in rank order in publicly accessible report cards or league tables (Cowley & Easton, 2004; Goldstein & Spielgelhalter, 1996; Morrison & Cowan, 1996; Zehr, 2001).

Reports on ranking schools have been produced in GB (Goldstein & Spiegelhalter, 2001; Morrison & Cowan, 1996; Thomas, 2001), in the US (Coe & Fitz-Gibbon, 1998; Lockwood et al., 2002; Zehr, 2001), in Canada (Cowley & Easton, 2004) and elsewhere (Wong, 1996). In GB, The Sunday Times State Schools Book was published as an annual table of national rankings for over 500 state schools in GB, which included intake and background variables, a variety of outcome variables and some contextual information (Morrison & Cowan, 1996). In his study, Wong (1996) ranked 27 Hong Kong secondary schools, but provided no rationale for why he did so. In the state of Indiana, Zehr (2001) explained that based on a set of performance reports, schools will begin to be placed in specific performance categories starting in the 2005-06 academic year. In the following year high-performing schools will be eligible for an award, while low-performing schools will be eligible for technical assistance aimed at improvement, including a change in personnel. Under the guidelines of the accountability plan the State could take over a school if it has demonstrated low improvement for five consecutive years (Zehr, 2001).

In Canada, Cowley and Easton (2004) of the Fraser Institute published a study entitled, "Report Card on British Columbia's Secondary Schools: 2004 Edition." The foundation of the report card was an overall ranking of each school's academic performance published annually by the Fraser Institute. Each school was rated on an overall scale from zero to ten. This rating was based on a standardized score calculated from each school's academic performance on eight selected indicators (these indicators were discussed in Chapter 1).

With the values obtained from these indicators, the authors blended public and independent secondary schools in BC into their analysis using only aggregated data. The instigators of these types of school-to-school comparison reports professed that they provide "valid and reliable information" to parents and students in making better decisions when selecting schools (Cowley & Easton, 2004; Morrison & Cowan, 1996). However, the validity of inferences made from such reports is questionable, as the indicators for outcome and contextual factors by which the institutions measure effectiveness varies greatly (Chester, 2003; Goldstein & Spielgelhalter, 1996; Schafer, 2003; Linn, 2001), and there is no single interpretation of the concept of school effectiveness (Cistone & Bashford, 2002; Coe & Fitz-Gibbon, 1998).

Summary of the Literature Review

The validity of interpreting the true effectiveness of a school depends on the choice of indicators measured (Coe & Fitz-Gibbon, 1998; Gray et al., 1995; Messick, 1995). Even where available measures are judged to be acceptable, there are inevitable limitations in making direct comparisons between institutions (Gibson & Asthana, 1998; Goldstein, 1997; Goldstein et al., 1992; Gray et al., 1995; Thomas, 2001; Yang & Goldstein, 1999; Waxman & Huang, 1997). From the literature, it is clear that there was no single dimension along which schools differed (Linn, 2001; Lockwood et al., 2002) and that the generalizability of the results from performance indicators on effective or ineffective schools was limited (Cistone & Bashford, 2002; Cronback & Linn, 1997; Lockwood et al., 2002). Even less information could be obtained from league tables or report cards that ranked schools. The rank-ordered table produced by Cowley and Easton (2004) was only constructed from a single composite score. It is not meaningful to evaluate entire

institutions with a single numerical score. If intake and contextual effects are ignored, confidence intervals around estimates of school effects will be too large to rank schools with any precision (Goldstein & Rasbash, 1993). The difference of a few percentage points could make a significant difference to the rankings, but would not be statistically or practically significant. In other words, in a rank-ordered table of 100 schools, there may be little difference between the schools ranked #2 and #12, but the uninformed reader naturally thinks otherwise. Only the very highest from the very lowest ranking schools would be statistically distinguishable. The uncertainty attached to individual institutions' results, at least based upon a single year's data is such that fine distinctions and detailed rank orderings are statistically invalid (Goldstein et al., 1992; Goldstein & Rasbash, 1993; Goldstein & Speilgelhalter, 1996; Messick, 1995). At best, it is suggested in the literature that performance indicator reports should only be used as feedback to individual institutions about potential problems and successes but not for comparisons between institutions.

Those who develop league tables should accept the responsibility of validating the variables included in their analyses, and be liable for the consequences of their use (Cronbach & Linn, 1997; Goldstein & Spielgelhalter, 1996; Morrison & Cowan, 1996). Data to be judged must be accurate and independently verified, because as Goldstein & Spielgelhalter (1996) state, no type of analysis can overcome inappropriate or poor variable selection. In would appear that the general public perceived these rankings as important (Cowley & Easton, 2004; Morrison & Cowan, 1996). Cowley & Easton (2004) claimed that they would keep producing their report cards until there is no longer a demand for them. Therefore, it is important that a closer analysis of the "Report Card on British

Columbia's Secondary Schools" be investigated for its validity and statistical relevance. By applying a two-level multilevel model approach, the present study purports to determine if further information about the variation between secondary schools can be more appropriately explained by including additional student level variables and contextual school level variables.

The literature discussed in this Chapter acknowledged that effective schools have been identified using many different statistical methodologies over the past 40 years. Different modelling procedures have led to different conclusions about individual explanatory variables, and the level of analysis choice for investigating these variables. However, despite their disagreements the underlying goal of all school effectiveness researchers was to isolate the characteristics that distinguished effective schools from the rest. In this effort, there have been significant contributions and improvements made over the years to the school effectiveness paradigm. As was demonstrated in the discussion of criticisms to this research genre, strong feelings were ignited that illustrated how researchers either opposed or supported these types of investigations. While some critics had rather passionate comments against school effectiveness research, all contributed to the pedagogical discourse that aims to mature this field of study. One of the most impressive contributions made to school effectiveness research was the introduction in the 1980s of a powerful modelling technique called multilevel modelling. Most early school effectiveness initiatives poorly conceptualized the many ways in which contextual variables could impact upon the learning and performance at the student, classroom, and school levels. These models generally assumed the same effects across groups, which failed to explicitly incorporate the hierarchical nature of students clustered into schools. Whereas multilevel

modelling simultaneously models this nested structure so that an investigation can be conducted on how student performance is influenced at the individual level as well as at the classroom and school levels. With the onset of computer programs sophisticated enough to handle these complex structures, the multilevel method increased with popularity. This has now become the preferred approach for investigating school effectiveness and improvement research. However, there continue to be differences in the perception of what school effectiveness is and how it should be measured and reported. There are even some researchers who have not adopted the multilevel method in their models and thus, the debate on which unit of analysis to study continues for them.

Chapter Three

Research Methodology

Research Design and Rationale

Chapter Two concluded by indicating that not all researchers interested in school effectiveness and improvement have implemented multilevel modelling techniques into their analyses. Furthermore, they have neglected to include relevant intake measures and school contextual variables as well. Some have purposively used student data aggregated to the school level, even though disaggregated information is usually more appropriate for most educational studies. Primarily for these reasons, many researchers have advised against using league tables because of the data limitations and inaccuracies of the statistical methods used for ranking schools. They are often compiled using only one year's worth of aggregated data, are usually not adjusted for intake measures, and rarely include many, if any, contextual variables. Yet as noted in the literature review, those types of rankings are still being published and in some cases, such as in Indiana, the results of these rankings are directing high-stakes, policy decision-making.

It is obvious that there is continued and escalating interest in measuring school effectiveness and improvement. In September 2003, the BC Ministry of Education formed a task force to look at improving student achievement across the province. The Student Achievement Task Force recommended a five-point action plan that included $180,000 for a school improvement strategy to be jointly developed with the Ministry of Education and The University of British Columbia. Sixty schools were included in the first year of the plan in 2003/04 and were eligible for monetary improvement awards in 2004/05 (Wickstrom et al., 2003). The BC Ministry of Education also assigned a task force to

investigate challenges and opportunities for education in rural areas and to investigate the reasons behind a decade of declining enrollments (Imrich et al., 2003). The Rural Task Force reported that about 15% of BC's public school students attend rural schools. Some of these are not just rural, but are considered remote and isolated with no road access, which adds to the challenge of providing quality education to these students. There are different challenges for rural communities compared with urban ones. The Task Force identified problems in recruiting and retaining teachers in some of the rural areas. Also, they acknowledged how many of BC's rural communities have economies that rely on one economic sector or one main employer. If something negatively influenced the industry or employer so that they were required to downsize their companies, often many employees would not find work in their communities and would be forced to leave. In support of these claims, Thomas (2001) also found that regional differences appeared to exist in the size and impact of school effects and encouraged additional researchers to include these types of contextual school variables for further analysis in their studies.

While the Fraser Institute's report card, prepared by Cowley and Easton, has not been used for educational policy decision-making in BC, the authors claimed it played an important role in the decision-making of parents when selecting schools (Cowley & Easton, 2004). Therefore, it is essential to statistically investigate the methods used by Cowley and Easton in terms of the validity of inferences made from their rankings. The foundation of the report card was an overall rating out of 10 of each secondary school's academic performance on eight selected indicators compiled from aggregated student information: (1) average provincial examination mark; (2) percentage of provincial examinations failed; (3) difference between school mark and examination mark in provincially examinable

courses; (4) difference between male and female students in the value of indicator #3, for English 12 only; (5) difference between male and female students in the value of indicator #3, for Mathematics 12 only; (6) provincially examinable courses taken per student; (7) graduate rate; and (8) composite drop-out rate. The authors suggested that these key academic indicators of secondary school performance provided the reader with measures of effective teaching, a valid assessment of academic counselling, and an accurate evaluation of school effectiveness and potential for improvement.

Cowley and Easton have been criticized on many aspects of their report card; not the least was with the lack of differentiation between independent and public schools in their model. It was argued by opponents to school rankings that there are considerable differences between these two sectors for school policies, governance, accountability, and school composition and climate. Public schools rely on school and district planning for the continued process of collaboration and coordination of services. While school boards that provide direction for school planning and the development of services generally govern independent schools. Studies have suggested there are differentiating effects of the distribution of academic achievement in schools from different sectors (Raudenbush & Bryk, 2002) and additional research implying that the regional location of schools affects academic achievement (Thomas, 2001). Therefore, the differential impact of independent and public schools as well as regional location of schools will be investigated in the present study. Furthermore, this current study will integrate disaggregated student level data and aggregated school level data into a two-level multilevel model to investigate how much of the variability in academic performance on provincial examinations is attributable to

differences between schools, and how much to differences between students within schools, which is something the current model used by the Fraser Institute is unable to do.

Using the software program, HLM 6.0 Hierarchical Linear and Nonlinear Modeling (Raudenbush, Byrk, & Congdon, 2000), to analyse the data as a two-level multilevel model this current study will explain the differentiating influence of level-two school characteristics on level-one student characteristic relationships, by building two intercepts- and slopes-as-outcomes models – Example A and Example B. Example A has as its criterion variable the average provincial examination percentage, which was the examination variable used in the Fraser Institute Rankings. Example B has the average English provincial examination percentage as its criterion variable. The within-school relationships are represented by the regression coefficients in the level-one models. The effects of school-level variables on each of these relationships are represented in the level-two models. The results from these models indicate whether residual slope variability remains after adding school contextual variables. In other words, the overall intercepts- and slopes-as-outcomes models will help to illustrate how differences among schools in their organizational characteristics and locations might influence the distribution of achievement within and among schools (Raudenbush & Bryk, 2002). This is a unique characteristic of multilevel modelling (Heck & Thomas, 2000). Another benefit of multilevel modelling is that the schools do not need to be balanced in numbers. School A does not have to have the same number of students in the sample as School C for the analysis. This is an important consideration for the current project because Grade 12 school enrollments vary considerably across schools in the sample. Multilevel modelling

provides a powerful framework for exploring how average relationships vary across hierarchical structures.

Study Sample

The school level data used in this study are from 256 secondary schools in BC are similar to those used in the Fraser Institute Rankings for the 2004 Edition. In addition to the school level data, the current study also incorporates student level information. Student level data was provided on behalf of the Ministry of Education by EduData Canada - an educational data provider housed at the University of British Columbia. Students in this current study were enrolled in Grade 12 during the academic year 2002/03 and wrote their applicable provincial examinations during this academic year. The present study involves schools of different sizes, locations (rural and urban), and includes public and independent schools throughout the Province of BC. Among the entire cohort of BC schools, only schools with more than 15 students with overall examination marks were included in Example A, and only schools with 10 or more students with English marks for Example B. These criteria for inclusion were determined because estimates of school effects based on small samples may not be reliable, making it difficult to interpret significant between-school differences (Mok, 1995). Schools that had missing data at level-2 were excluded because there can be no missing data at level-2. Also excluded from this current study were continuing or adult education centres as well as certain alternative schools that do not offer a full program of provincially examinable courses. Based on these criteria, the final study sample for Example A includes 43,146 grade 12 students in 254 secondary schools, and Example B includes 35,887 Grade 12 students in 251 secondary schools. These schools are located in 57 BC school districts and authorities (Francophone Education

Authority) across four regions of BC. The grade 12 sample sizes for Example A at each school range from 15 to 545, and for Example B range from 13 to 488.

Criterion Variables in the Study – Examples A and B

The average provincial examination mark is the criterion variable in Example A in this study. It is a composite score comprised of adding up the percentage earned in each provincially examinable subject area for all examinations written by a student, divided by the number of examinations written by each student within the 2002/2003 school year. In the case where one student wrote the same subject examination twice, the highest percentage was included in the calculation as per the 1995 Ministry Graduation Program. Each examination is weighted the same, so two students could achieve the same total score, but have achieved different scores for different examinations according to subject. For example, a student who scored 60, 75, and 70 on Principles of Mathematics 12, English 12 and Science 12 exams respectively, and a different student who scored 73, 68, and 64, would both earn the same composite score – 68.3. Examination marks are collected and recorded electronically by the BC Ministry of Education.

Cowley and Easton of The Fraser Institute reasoned in the 2004 Edition of the report card that the provincial examinations were designed to achieve a distribution of results reflecting the differences in students' mastery of the course work. They argued that differences among students in interests, abilities, motivation, and work habits would inevitably have some impact upon the final results. However, they suggested that there were recognizable differences from school to school within a district in the average results on the provincial examinations and there was also variation within schools in the results obtained in different subject areas. They argued that these differences in outcomes could

not be wholly explained by the individual and family characteristics of the school's students. Thus, they believed it was reasonable to include average examination mark for each school as one indicator of effective teaching. However, Goldstein & Rasbash (1993) and Goldstein & Thomas (1996) briefly discussed the differences of using individual subject areas as the criterion compared with using total examination scores. They argued that by using total examination scores, fine distinctions and detailed rank orderings were statistically invalid and that underlying relationships could be masked. By calculating the estimated residuals with a 95% confidence interval for each school effect, they demonstrated that there was considerable overlap of intervals, which suggested that it was not statistically possible to discriminate easily between schools based on total examination results. In the current study, two examples have been designed. One includes the average examination mark as suggested by Cowley and Easton, and the other includes the English examination mark to determine if additional information can be obtained about school variance based on a single subject area rather than an average overall score as the criterion (Goldstein & Rashbash, 1993; Goldstein & Thomas, 1996). The rationale for choosing the subject area, English 12, is that it had the highest provincial participation rate of all the provincially examinable courses for the sample of schools used in this study. Also, it is one of the two Grade 12 examination subjects the Ministry of Education publicly reports results on their website for School Profiles, because it is said to demonstrate knowledge in a key area of BC provincial curriculum – literacy. However, not all students are required to enroll in English 12. There is the option of enrolling in Communications 12, which is also a provincially examinable subject. Therefore, is it is important to keep in mind that the

personal characteristics of a student enrolled in English 12 may differ from a student enrolled in Communications 12.

Predictor Variables: Student Level (or Level-One)

Unlike the Fraser Institute's study, this current investigation includes student level-one variables that are nested in the higher-level school unit. Instead, Cowley and Easton (2004) used student results aggregated to the school level. Three predictor variables were included in level-one of the model for both examples: (1) school mark, (2) gender, and (3) ESL. Further information on each of these variables is provided below.

School Mark is a percentage earned by the student on the accumulation of all the results from in-class assessments for the same provincially examinable courses included in the study. In Example A, school mark is the mark earned by each student for the same courses included in calculating the provincial examination average mark. In Example B, the school mark is the percentage achieved in the English 12 school course. Cowley and Easton (2004) suggested using an indicator of grading gap that was the difference between the school mark and provincial examination mark to determine whether or not a school was providing an accurate estimate of the extent to which knowledge of the provincial curriculum was being acquired or if a school was consistently granting marks substantially higher or lower than those achieved on the provincial examinations. The grading gap indicator used in the Fraser Institute study can be problematic because the difference between the grades could be attributed to regression-to-the-mean. Regression-to-the-mean refers to the phenomenon that subjects who deviate markedly from the mean (either high or low scorers), when retested tend to regress or score closer to the group mean (Hopkins & Glass, 1996). It appears that the Fraser Institute report did not take this phenomenon into

account when determining the grading gap, because Cowley and Easton offered no discussion on how they interpreted the difference between these marks other than attributing it to the teachers, when clearly some of these differences need to be attributed to the regression effect. As was suggested by Bobko (2001), by having a sample of students' with a complete range of scores, the regression effects would hopefully average out (some regression up to the mean, and some down to the mean). However, this would require testing on each sample under investigation. Another concern of the Fraser Institutes study was that they used absolute values in their tables for grading gaps and positive or negative differences cannot be determined by using absolute values. School marks will be included in the current study as a predictor.

Gender is a dichotomous variable coded, 1=male and 0=female. Cowley and Easton (2004) identified research that has found systematic sex-based differences in academic results in BC's secondary schools. They suggested that these differences are particularly apparent where the local school rather than the Ministry of Education designs assessments. It is then important to include gender as a variable when investigating achievement.

ESL Status is included in this study as a background characteristic of the student. This is a self-defined characteristic, and is a dummy-coded variable where 1=ESL and 0=non-ESL. It is important to determine if ESL status has a differentiating effect on achievement.

Location of Level-One Predictors

The choice on where to locate these level-one predictor variables is very important. In the simple model, the intercept β_{oj} is defined as the expected outcome for a student attending school j who has a value of zero on X_{ij}. If this is not meaningful, then the researcher can transform X_{ij} to make the intercept β_{oj} more meaningful by group-mean

centering ($X_{1ij} - \overline{X}_{1\bullet j}$), grand-mean centering ($X_{1ij} - \overline{X}_{1\bullet\bullet}$), or locating it on another metric that makes sense to the researcher (Raudenbush & Bryk, 2002). In general, if the mean of X_{ij} systematically varies across schools, the choice of centering impacts the inferences made about $\hat{\tau}_{00}$. Raudenbush and Bryk (2002) recommend group-mean centering, also known as centering within context, to detect and estimate the slope heterogeneity across schools. When variables are group-mean centred, the interpretation of the within-group slopes is the expected outcome for a student whose value is equal to the school average on all predictors. Alternately, when variables are grand-mean centred, the intercept is the expected outcome for a student whose value on the predictor is equal to the grand mean of the total sample.

Even dichotomous variables such as gender can be centred. In this current study, X_{ij} takes on a value of 1 if subject j is a male and 0 if female, so if β_{oj} is not centred it is the expected outcome for a female student in school j (i.e., the predicted outcome for a student where $X_{ij} = 0$). However, if gender is group-mean centred then for males, $X_{1ij} - \overline{X}_{1\bullet j}$ will take on the value equal to the proportion of female students in school j; for females it will take on the value equal to minus the proportion of male students in school j (Raudenbush & Bryk, 2002). By group-centering the variable gender makes the score of males in a female dominated school different from scores of males in a dominant male school. Thus, males in one school no longer have the same score as males in another school, but instead a score that deviates from the class percentage of gender.

When group-mean centering is applied to level-one predictors, the cross-level interaction and between-group interaction can be differentiated (Hofmann & Gavin, 1998).

There is no statistically correct choice for centering level-one variables (Davison et. al., 2002; Hofmann & Gavin, 1998; Kreft, de Leeuw, & Aiken, 1995). However, the choice of location should be couched in theory. For this current study, the relationship of interest is how attending a particular school influences students' achievement on the overall provincial examination (Example A) and the English 12 provincial examination (Example B). Using Kreft, de Leeuw, and Aiken's analogy of the frog-pond effect, each student's achievement is predicted from being considered a big or small frog in that particular pond, as well as taking into account the pond's size, type, and location in relationship to the other ponds in the sample. Therefore, success or failure on these outcomes is considered, in part, a school characteristic and for these reasons all level-one variables are group-mean centred and all level-two variables are grand-mean centered in this present study (Kreft, de Leeuw, and Aiken, 1995).

Predictor Variables: School Level (or Level-Two)

In this study schools are considered level-two variables that typically explain differences in school effectiveness as a result of incorporating school level variables into the analyses, while controlling for the level-one within-school variables. This current study includes eight school-level variables in the analyses: 1) average number of examinations written, 2) graduation rate, 3) proportion of ESL, 4) proportion of Aboriginal students, 5) proportion of Male students, 6) average years of parents' education, 7) school sector, and 8) region (over three vectors). As previously explained all these level-two variables will be grand-mean centred. A more detailed explanation of how these variables are derived is provided below.

Average Number of Examinations Written is the number of provincial examinations written by each student in a particular school divided by the number of students participating in the provincial exams within the sample.

Graduation Rate indicator compares the number of potential graduates enrolled in the school on September 20 with the number of students who actually graduate by the end of the same school year. Only those enrollees who were capable of graduating with their class within the current school year were included in the count of potential graduates. This indicator varies widely between schools throughout the province. It was collected from the BC Ministry of Education in one of their standard reports compiled annually for each school.

Proportion ESL is included as a characteristic of the school for the 2002/03 year. Students identified themselves as either ESL or non-ESL for this particular year. A student may have identified him/herself as ESL last year, and non-ESL this year. This self-declared variable was dummy-coded at the student level where 1=ESL and 0=non-ESL. The student file was then aggregated to the school level, which involved summing the 1's for each school and taking this total as a percentage of the number of students per school included in the study. This is a proportion of ESL students per school, enrolled in Grade 12 included in the present study's sample rather than a proportion of ESL students that may actually be enrolled in Grade 12 in the school. There could be ESL students enrolled in Grade 12 in a school, who did not write a provincial examination within the 2002/03 academic year and so were not included in the calculation of the proportion ESL.

Proportion Aboriginal is included in this study as another characteristic of the school. There are areas in BC where the concentration of Aboriginal students is high and where

some school districts have implemented education programs designed specifically to increase the academic achievement of aboriginal students. It is important to see if the impacts by district are amplified for aboriginal students. This self-declared variable was dummy-coded at the student level where 1=aboriginal and 0=non-aboriginal. Again, the student file was aggregated to the school level, which involved summing the 1's for each school and taking this total as a percentage of the number of students per school included in the study. This is a proportion of Aboriginal students per school enrolled in Grade 12, included in the present study's sample rather than a proportion of Aboriginal students that may actually be enrolled in Grade 12 in the school.

Average years of parents' education was the variable used by Cowley and Easton (2004) because it was found in their multiple regression analyses to be the only home characteristic defined that was statistically significantly related to overall exam achievement. This indicator was not used in the calculation of the overall school rank by the Fraser Institute, but was used in the post analysis and the results reported in the published rank tables as an adjusted rating. Using enrollment data from the Ministry of Education, sorted by Dissemination Area and linked to home variables from 2001 Census data provided by Statistics Canada, they established a profile of the student body's home characteristics for each school. The current study incorporates this average years of parents' education indicator calculated by the Fraser Institute as one of the school level explanatory variables.

The School Sector indicator identifies the school to be either independent or public (dummy coded where 1=independent and 0=public). Independent schools are defined by the Ministry as any person or organization outside the public school system providing a

program of education to 10 or more school age students. Data was provided only for those independent schools that had a teaching staff that was at least 80% certified and offered enough provincially examinable courses to meet the inclusion criteria of the current study. Independent schools can be categorized differently according to the level of supplemental funding received by the provincial government. For the purposes of this study, these categories have not been included. Public schools are defined as educational facilities funded 100% by the provincial government. Studies on school effectiveness have found significant differences when investigating the differential effectiveness across school sectors (Raudenbush & Bryk, 2002). The report card includes both public and independent schools in the analysis, but has not tested this relationship with academic achievement on provincial examinations. Therefore, the current study will investigate this relationship.

The Region Vectors identify which of the four regions in BC the school is located; the Lower Mainland, Vancouver Island and the Pacific Coast, Fraser Valley and Southern BC, and the Interior and Northern BC. These are the same regional boundaries suggested by Cowley and Easton in the 2004 Edition of the Report Card. They are categorical variables that require dummy coding for the analysis. With multiple categorical variables, the number of coded vectors generated is equal to the number of categories minus one (or degrees of freedom). Therefore, for Region there will be three coded vectors: R_1, R_2, and R_3. If a school is located in the Lower Mainland, it will be coded 0 in all three vectors (R_1, R_2, and R_3), if a school is located in the Vancouver Island and the Coast Region it will be coded 1 in R_1 and 0 in vectors R_2 and R_3, schools in the Fraser Valley and Southern BC would be coded 1 in R_2 and 0 in R_1 and R_3 vectors, and if located in the Interior and Northern BC the school would be coded 1 in R_3 and 0 in the other two vectors. The results

from the analyses with respect to Region (R_1, R_2, and R_3) will be compared to the expected results for the schools located in the Lower Mainland Region.

Data Collection

Grade 12 provincial examinations are part of the BC graduation requirement. The Ministry claims that these exams ensure fair treatment when applying for post-secondary education, and provide opportunities to compete for provincial scholarships. Students generally write their provincial exams at the end of their grade 12 school year in June, but they can also write in November, January, and April. All of these data are available for all schools because they are required standardized exams, and as a result of the Ministry's initiatives to provide accountability measures. Results from these examinations for all students attending all schools in BC are collected and recorded by the Ministry of Education. The data for the current study, including student and some school data for the 2002/03 academic year, were requested from the Ministry of Education through EduData Canada. Research Analysts at EduData linked records on student personal characteristics, school marks, schools attended, and provincial examination results using Personal Education Numbers (PENs) that are assigned to all students in the K-12 education system in BC by the Ministry of Education. The school data set was supplied directly from the Ministry of Education as an exact cut of the data requested by the Fraser Institute. All data were provided by the Research Analyst at EduData Canada in SPSS format on two CD-Rom disks. The data were then re-organized using Microsoft ACCESS, a relational database tool, into a format that was more accessible for analysis. As mentioned previously, the school variable, average years of parents' education, was collected from the

published version of the Fraser Institute rankings and manually added to the school level data file.

Model Design: A Two-Level Intercepts- And Slopes-as-Outcomes Model

The first level of the model (student model) examines the relationships between overall academic achievement on provincial examinations and four parameters: an intercept and three regression coefficients (gender, school mark, and ESL status).

Level-One:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{1ij} - \overline{X}_{1\bullet j}) + \beta_{2j}(X_{2ij} - \overline{X}_{2\bullet j}) + \beta_{3j}(X_{3ij} - \overline{X}_{3\bullet j}) + r_{ij}. \quad (1)$$

Or more succinctly,

$$Y_{ij} = \beta_{0j} + \sum_{j=1}^{3} \beta_{qj}(X_{qij} - \overline{X}_{q\bullet j}) + r_{ij}. \quad (2)$$

q=1, 2, 3.

Where Y_{ij} is student i's achievement in school j, which is assumed to be normally distributed $y \sim N(XB, \Omega)$;

β_{0j} is the mean academic achievement on all provincial exams written in school j;

β_{1j} is the average effect of school mark on average provincial exam achievement in school j;

β_{2j} is the average effect of gender status on achievement for school j (i.e., the "gender achievement gap" – the mean difference between achievement of male and female students); and

β_{3j} is the average effect of ESL status on achievement for school j (i.e., the "ESL achievement gap" – mean difference between scores of ESL and non-ESL students).

The individual variance, σ^2, represents the residual variance at level-one, r_{ij}, that remains unexplained after taking into account the three regression coefficients. Each of these distributive effects ($\beta_{0j}, \beta_{1j}, \beta_{2j}, \beta_{3j}$), are net of the others. For example when investigating the influence of gender, the other two variables are controlled. Within the level-one model, each school can have a different average achievement (i.e., intercept) and a different impact of three variables on average academic achievement (i.e., slope). In terms of this model, an effective school would be characterised by a high level of mean academic achievement (i.e., a large positive value for β_{0j}), small gender and ESL achievement gaps (i.e., a small negative value for β_{2j} and β_{3j}), and weak differentiating effects for the school mark variable (i.e., small positive value for β_{1j}). The second level of the model (school model) examines the effects of the eight school level variables on level-one relationships. Of the level-one coefficients, gender and ESL status are considered fixed, while school mark is specified as random in the level-two model.

$$\text{Level-Two: } \beta_{qj} = \gamma_{q0} + \sum_{j=1}^{10} \gamma_{qk}(W_{kj} - \overline{W}_{k\bullet}) + u_{qj}. \qquad (3)$$

q=0, 1, 2, 3. k=1, 2, 3, …, 10.

Where,

W_{1j} = average number of examinations;

W_{2j} = graduation rate;

W_{3j} = proportion ESL;

W_{4j} = proportion Aboriginal;

W_{5j} = proportion male;

W_{6j} = average years of parents' education;

W_{7j} = school sector (1=independent, 0=public);

W_{8j} = in which region the school is located: R_1 (coded 1 if school belongs to vector R_1=Vancouver Island and the Coast)

W_{9j} = in which region the school is located: R_2 (coded 1 if school belongs to vector R_2=Fraser Valley and Southern BC)

W_{10j} = in which region the school is located: R_3 (coded 1 if school belongs to vector R_3=Interior and Northern BC)

$\gamma_{00},\ldots,\gamma_{010}$ = level-2 intercept/slopes to model β_{0j}

$\gamma_{10},\ldots,\gamma_{110}$ = level-2 intercept/slopes to model β_{1j}

$\gamma_{20},\ldots,\gamma_{210}$ = level-2 intercept/slopes to model β_{2j}

$\gamma_{30},\ldots,\gamma_{310}$ = level-2 intercept/slopes to model β_{3j}

$u_{0j},\ldots,u_{1j}$ = are level-2 random effects (gender and ESL are level-2 fixed effects)

Because there are only two level-2 random effects, the variances and covariance among them form a 2 x 2 matrix.

$$T = \begin{bmatrix} \tau_{00} \\ \tau_{10}\tau_{11} \end{bmatrix}$$

Finally, the combined model looks like this:

$$Y_{ij} = \gamma_{00} + \sum_{j=1}^{3} \gamma_{q0}(X_{qij} - \overline{X}_{q\bullet j}) + \sum_{j=1}^{10} \gamma_{0q}(W_{kj} - \overline{W}_{k\bullet})$$

$$+ \sum_{j=1}^{30} \gamma_{qk}(X_{qij} - \overline{X}_{q\bullet j})(W_{kj} - \overline{W}_{k\bullet}) + \sum u_{1j}(X_{qij} - \overline{X}_{q\bullet j}) + u_{oj} + r_{ij}. \quad (4)$$

The part of the equation, $\sum_{j=1}^{30} \gamma_{qk}(X_{qij} - \overline{X}_{q\bullet j})(W_{kj} - \overline{W}_{k\bullet})$, represents the cross-level interaction between level-one X_{qij} (group-mean centred) and level-two W_{kj} (grand-mean centred) variables. These can be understood as the moderating effects of level-two variables on the relationships between level-one predictors X_{qij} and the outcome, Y_{ij}. Also note that the error term has become more complicated, $u_{1j}(X_{1ij} - \overline{X}_{q\bullet j}) + u_{0j} + r_{ij}$. This accommodates the relationship between u_{0j} and u_{1j}, which are common to every level-one observation within each level-two unit. In this current study only school mark is considered a random effect at level-two, while ESL status and gender are considered fixed. This is because there are some schools included in the sample in which these two variables are homogeneous and do not vary.

Model Assumptions

Prior to applying a multilevel model, the theoretical assumptions of such a model require consideration. The efficiency and power of this multilevel approach rests on pooled data across the units comprising of two-levels, which implies the requirement of a large

dataset. Mok (1995) provides a detailed simulation analysis of sample size requirements for two-level designs in educational research. With less than adequate power there is an unacceptable risk of not detecting cross-level interactions. However, it is generally accepted that there should be adequate statistical power with 30 groups of 30 observations each, or even 150 groups each with 5 observations. This current analysis should meet the sample size requirement with 254 groups (schools) of at least 15 observations (students) each in Example A and at least 10 observations for each of the 251 schools in Example B. Another assumption is that at level-one, errors are normally distributed and are homogeneous, that is, Var(r_{ij},) = σ^2. Raudenbush and Bryk (2002) suggested that statistical evidence recommends that the estimation of the fixed effects, γ, and their standard errors will be robust to violations of this assumption. However, there are some situations where the heterogeneity at level-one may require modelling. Also for level-one, under the null hypothesis, it is expected that the average achievement for school j will equal the average school mean for all j schools, and the slopes of school j will equal the average of the slopes for all j schools. At level-two, the tau's (τ) are the variances of the intercepts and slopes, and the covariance between them, and it is assumed that the school-level residuals follow a multivariate normal distribution with variances (τ_{00}, τ_{11}) and covariance (τ_{10}). This dependency violates the assumption in ordinary regression of independent errors across observations, but can be handled using a multilevel approach (Heck & Thomas, 2000). Another consideration of multilevel modelling is that missing data at level-one can be handled, but there cannot be missing data at level-two (Raudenbush & Bryk, 2002).

Ranking Schools with Empirical Bayes Estimators

After computing school estimates, researchers can rank these estimates to identify the most effective and least effective schools (Pituch, 1999). First, the goal is to find the best estimator of $\hat{\beta}_{qj}$. In order to increase the accuracy of estimating β_{qj} in an intercepts- and slopes-as-outcomes model, empirical Bayes estimators can be computed that shrink the estimates toward predicted values of β_{qj}. Empirical Bayes estimates are more beneficial than OLS regression or ANCOVA, because unlike OLS it takes into account group membership even when the number of groups (i.e., schools) are large, and produces relatively stable estimates even when sample sizes per school are small or moderate (Raudenbush & Bryk, 2002). ANCOVA does take group membership into consideration, but this tends to be impractical when the number of schools in the sample is large. Using Equation 5, the empirical Bayes point estimates can be calculated as

$$\beta^{*}_{qj} = \Lambda_j \hat{\beta}_{qj} + (\mathrm{I} - \Lambda_j) W_{kj} \hat{\gamma}_{qk} . \tag{5}$$

Where, $\Lambda_j = \mathrm{T}(\mathrm{T} + V_j)^{-1}$ is the ratio of the parameter dispersion matrix for β_{qj} (i.e., T) relative to the total dispersion matrix for the $\hat{\beta}_{qj}$, which contains error and parameter distribution (e.g., $\mathrm{T}+\mathrm{V_j}$). Raudenbush and Bryk (2002) suggested that, Λ_j could be considered a multivariate reliability matrix. When empirical Bayes estimators are used in a multilevel model they can provide stable indicator for judging individual school performance (Raudenbush & Bryk, 2002). Researchers also calculate the empirical Bayes residual estimates for ranking schools. This type of model hypothesizes that all students within school j have an effect, u_{qj}, added to their expected score as a result of attending

that school. The formula for calculating the empirical Bayes residual estimates in an intercepts- and slopes-as-outcomes model is shown in Equation 6.

$$\text{Estimating empirical Bayes residuals: } u^{*}_{qj} = \beta^{*}_{qj} - [W_{kj}\hat{\gamma}_{qk}]. \quad (6)$$

The empirical Bayes point estimates and 95% confidence intervals will be calculated in this current study and the secondary schools and the estimates from both Example A and B will be discussed.

Benefits of Applying an Intercepts- and Slopes-As-Outcomes Model

By applying the intercepts- and slopes-as-outcomes multilevel technique, the current investigation aims to clarify and extend previous research concerning the definition of measurement of secondary school effectiveness in BC across a range of regional contexts. The objective of this analysis is to establish a multilevel model for measuring school effectiveness in terms of student outcomes on provincial examinations and English in Grade 12 that take into consideration the educational context of schools located in different sectors and regions. It will further incorporate the disaggregated level of information (student level) into the analyses. Three aspects of multilevel analysis will be examined: (1) the statistical significance of different explanatory variables included in the model, (2) the percentage reduction in the total and school level variation in student outcomes by introducing different explanatory variables in the model, and (3) the percentage of total variation attributable to the student and attributable to the school. Chapter Four will provide the research findings, as well as a discussion on how valid report cards and league tables are for guiding the choice of an institution based on overall and English provincial examination marks.

Chapter Four

Research Results

This chapter reports the results of the two examples using multilevel intercepts- and slopes-as-outcomes models for investigating the variability in school performance on overall provincial examinations (Example A) and English 12 examinations (Example B) for students enrolled in a sample of BC schools during the academic year 2002/03. The methodological significance of this study is that it demonstrates the power of multilevel analysis to educational research and provides a much stronger theoretical framework for reviewing how schools differ compared to traditional linear methods. This model allows the researcher to partition the variance to determine how much could be attributed to differences between schools, and how much to differences between students within schools. The purpose of this current study is to take into consideration student characteristics such as students' overall school marks (Example A), students' English school marks (Example B), ESL status, and gender. In addition, contextual school effects are included such as average number of examinations written per school, school graduation rate, proportion ESL, proportion Aboriginal, proportion male, average years of parents' education, school sector (public compared with independent) and region (four regions within BC). A discussion on the inferences made from these findings compared to the assumptions made by Cowley and Easton (2004) in compiling the 2004 edition of the Report Card on BC Secondary Schools will be provided in the discussion section at the end of this chapter.

Results for Example A – Overall Provincial Exam mark as Criterion

The descriptive statistics for Example A level-one and level-two variables are described in Table 1. Prior to building a multilevel model, the data were analysed first to determine if the assumption of normality in the criterion and school mark predictor was met, and then to determine if they justified a two-level multilevel analysis. The initial step involved plotting the student data for the criterion and the school mark predictor along P-P plots using the Statistical Package for the Social Sciences (SPSS) 12.0 software. If these data were normally distributed they would follow a straight-line as the observed cumulative probability would be very close to the expected cumulative probability. Figures 1 and 2 demonstrate that the data for both of these variables in Example A were normally distributed. The next steps for determining whether the data could be appropriately analysed using a multi-level model included fitting the data to 1) a one-way ANOVA with random effects, 2) a random-coefficient regression model, and 3) an intercepts- and slopes-as-outcomes model. The results of each application for Example A are provided and discussed throughout the first section of this chapter, and the second section reports the results of each application for Example B.

One-Way ANOVA With Random Effects Model

The criterion variable for a one-way ANOVA with random effects model was fitted to the data for Example A, where the overall provincial examination mark was the criterion, and the residuals from each level are considered random. The results from the one-way ANOVA model with random effects provide information in order to determine the total amount of variability on the overall provincial examination scores achieved within and between schools. The average school mean, γ_{00}, was estimated as 68.17. The pooled

Table 1

Descriptive Statistics for Example A with Average Provincial Exam Score as Criterion

Variable Name	Mean	SD
Level-1 (n=43,146)		
Provincial exam score	68.36	12.03
School mark	71.13	12.55
ESL (0=Non-ESL, 1=ESL)	0.03	0.16
Gender (0=Female, 1=Male)	0.49	0.50
Level-2 (n=254)		
Average provincial exam score	68.09	4.03
Sector (0=public, 1=independent)	0.15	0.36
R1 (1=Vancouver Island and the Coast)	0.21	0.41
R2 (1=Fraser Valley and Southern BC)	0.23	0.42
R3 (1=Interior and Northern BC)	0.19	0.39
School Graduation rate	94.78	3.84
Average years of parents' education	14.44	1.15
Average number of exams written	2.87	0.57
Proportion Aboriginal	6.27	13.06
Proportion ESL	2.23	5.66
Proportion Male	49.00	8.98

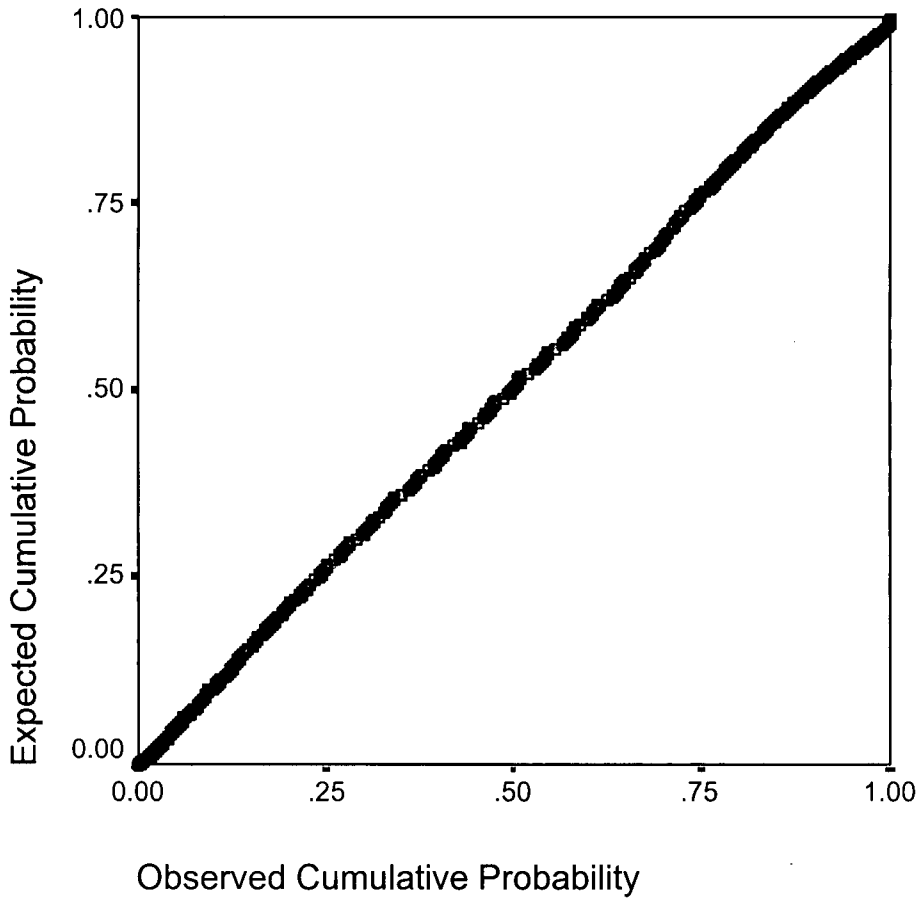

Figure 1. P-P Plot in SPSS to Determine Whether the Distribution of Student Level Average Overall Examination Mark is Normal in Example A.

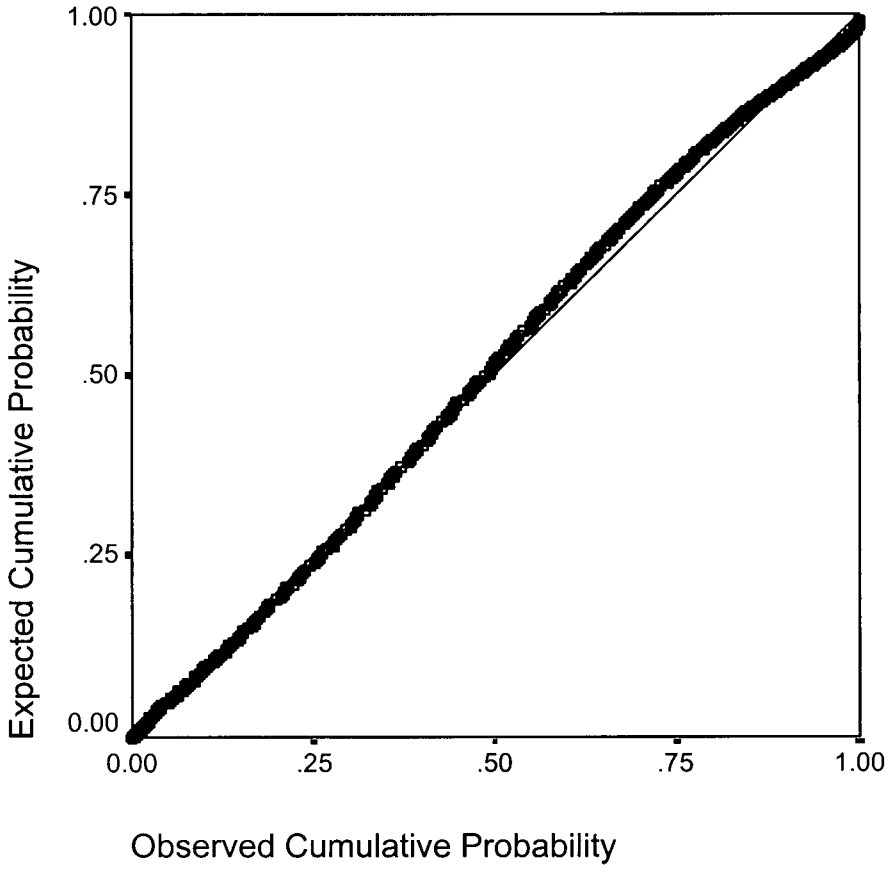

Figure 2. P-P Plot in SPSS to Determine Whether the Distribution of the Student-Level Predictor, Student Overall School Mark, is Normal in Example A.

within-school or level-1 variance, $\hat{\sigma}^2$, was estimated at 135.41, and the variance among the 254 school means, $\hat{\tau}_{00}$, was 12.48. Using Equation 7, the proportion of variance (or intra-class correlation) between schools can be determined.

$$\text{Proportion of Variance Explained} = \hat{\tau}_{00} / (\hat{\tau}_{00} + \hat{\sigma}^2). \quad (7)$$

Therefore, approximately 8.4% of variance in the outcome can be attributed to differences between the schools and the remaining 91.6% to differences between students within schools. It appears that most of the variance in the outcome is explained by within-school variability rather than by attending a particular school. However, using the HLM output, the 95% plausible values can be calculated by using Equation 8.

$$95\% \text{ Plausible Values} = \hat{\gamma}_{00} \pm 1.96(\hat{\tau}_{00})^{1/2}. \quad (8)$$

Even though there was only approximately 8.4% of the variance in average overall provincial examination scores that appear to be attributed to the school, the above equation provided information that the 95% plausible values had a wide range of about 13.84 points on average provincial achievement between schools (61.25, 75.09). Furthermore, the reliability of school sample means for this study was $\lambda = .90$, which implies for the schools in this sample there was considerable variability between schools on the outcome. By adding additional student and school-level variables the aim is to reduce this amount of unexplained variance between schools.

Random-Coefficient Regression Model

The next model designed was the random-coefficient regression model to represent the distribution of overall provincial achievement in each of the 254 schools. Specifically at the student level, the overall exam mark for student i in school j (Y_{ij}) was regressed on

overall school mark, ESL status, and Gender. Each school's distribution of achievement was explained in terms of four parameters: an intercept and three regression coefficients. As was described in Chapter 3, the school mark, ESL status, and Gender variables were all group-mean centred for this model. The results from the regression analysis are reported in Table 2.

In the present study, less than half (n=117) of the total 254 schools had variation on ESL status and 250 schools varied on gender status. Upon further investigation of the data, the schools that did not vary on gender and ESL status included a high number of independent schools. In this particular study, the impact of sector on school achievement was one of the main relationships to be examined. Therefore, these two variables were considered fixed in the model so that all data available could be used for the analyses.

In the random-coefficient regression model, the average school mean achievement on overall provincial examinations (γ_{00}) was estimated as 68.13. The level-1 variance was reduced from 135.41 in the one-way ANOVA with random effects to 70.73, after taking into account these three student-level variables. The average overall school mark differentiation ($\hat{\gamma}_{10}$) was positively related to the average overall exam mark .64. Gender ($\hat{\gamma}_{30}$) was also positively related to the outcome. This implies that in the average school, male students with similar ESL status and school mark scored 1.48 points higher on the provincial examination compared with their female schoolmates in Grade 12. In contrast, the ESL status, ($\hat{\gamma}_{20}$) was –4.56 points. This suggests that in a typical school, ESL students were scoring 4.56 points lower than non-ESL students with gender and

school marks like their own. The reported t ratios were quite large, indicating that each of the level-one predictors was statistically significant.

Table 2

Results from the Random Coefficient Regression for Example A

Fixed Effects	Coefficient	SE	t-ratio	p-value
Intercept, γ_{00}	68.13	0.24	281.18	0.000
Schlmark, γ_{01}	0.64	0.01	55.47	0.000
ESL, γ_{02}	-4.56	0.26	-17.79	0.000
Gender, γ_{03}	1.48	0.08	17.68	0.000
Random Effects	**Variance Component**	**df**	**χ^2**	**p value**
School mean, μ_{oj}	14.12	253	6156.68	0.000
Schlmark Slope, μ_{1j}	0.17	0.03	2168.15	0.000
Level-1 effect, r_{ij}	70.73			
Reliability of OLS Regression-Coefficient Estimates				
School mean achievement	0.95			
School mark differentiation	0.83			

The proportion of variance in level-one explained by this model can be calculated using Equation 9.

$$\text{Proportion of variance explained} = \frac{\hat{\sigma}^2(ANOVA) - \hat{\sigma}^2(REGRESSION)}{\hat{\sigma}^2(ANOVA)}. \qquad (9)$$

Applying equation 9, the proportion of variance explained for level-one in this model was 47.8%. The estimated level-two variances for the random-coefficient regression model provide empirical evidence about the variability in the relationship between exam marks and student level variables across schools. The homogeneity of variance test for the

level-two random effect can be used to test whether the relationship between overall school marks and overall provincial examination marks differ across schools. If not, it may be reasonable to assume that all schools have the same differentiating effect of school mark on the mean outcome. To test formally whether the estimated value of τ_{10} was significantly greater than zero, the hypothesis is $H_0 = \tau_{00} = 0$. In this case, the χ^2 test statistic took on a value of 2,168 with 253 degrees of freedom. The null hypothesis was highly implausible ($p < .001$), indicating significant variation did exist among schools in the differentiating effect of school marks on the achievement on overall provincial examinations. Plausible value estimates from the random coefficient regression model provide useful descriptive statistics of how much schools vary in terms of average provincial exam achievement and school mark differentiation effects. To gauge the magnitude of the variation among schools in their mean achievement levels, it is useful to calculate the plausible value range for these means. Under the normality assumption, it is expected that 95% of the school means fall within the range:

$$\hat{\gamma}_{q0} \pm 1.96(\hat{\tau}_{qq})^{1/2}. \qquad (10)$$

Thus, school means (β_{0j}) would be expected in the range of (60.76, 75.50) and school mark differentiation effect of (.31, .91). These results imply considerable variation, even more so than was demonstrated in the one-way ANOVA with random effects model. Contrary to what might be expected, it appeared that all schools engage in some kind of school mark differentiation in that values of β_{1j} were not plausibly zero.

The correlation among school effects can be calculated using Equation 11.

$$\hat{\rho}(\mu_{qj}, \mu_{q'j}) = \hat{\tau}_{qq'}/(\hat{\tau}_{qq}\hat{\tau}_{q'q'})^{1/2}. \quad (11)$$

Schools displaying high levels of achievement on overall provincial examination tended to be more differentiating with regard to school mark ($\hat{\rho} = 0.565$) than schools with lower achievement levels. Table 2 also reports the reliability for the level-two random effects. The reliability estimates suggested that there was considerable power in this dataset for examining the hypothesis about the effects of school characteristics on average provincial exam achievement, since the intercept estimates were highly reliable ($\lambda = .946$). Also, the reliability ($\lambda = .826$) indicates that these data were very useful for studying how school characteristics influence the academic differentiation of overall school mark.

Intercepts- and Slopes-as-Outcomes Model

Now that the variability of the regression equations across schools has been estimated, it is important to build an explanatory model to account for this variability. As described in Chapter 3, the Intercepts- and Slopes-as-Outcomes Model can provide information about why some schools have higher means than others on the overall provincial examination, and why in some schools the association between school mark and overall exam mark was stronger than in others. The results from the random coefficient regression model indicated that each of the level-one predictors had a statistically significant relationship with provincial examination scores. Further, there was statistical evidence provided by the τ_{10} point estimates, the χ^2 homogeneity test, and the reliability statistics (λ) to indicate that there was sufficient reliability among schools in the school mark differentiating effect to consider this random in the intercepts- and

slopes-as-outcomes model. All ten level-two variables as defined in Chapter Three were fitted to this model in the preliminary analysis as indicated in the combined equation 10.

$$Y_{ij} = \gamma_{00} + \sum_{j=1}^{3} \gamma_{q0}(X_{qij} - \overline{X}_{q\bullet j}) + \sum_{j=1}^{10} \gamma_{0q}(W_{kj} - \overline{W}_{k\bullet}) \qquad (10)$$

$$+ \sum_{j=1}^{30} \gamma_{qk}(X_{qij} - \overline{X}_{q\bullet j})(W_{kj} - \overline{W}_{k\bullet}) + \sum u_{1j}(X_{qij} - \overline{X}_{q\bullet j}) + u_{oj} + r_{ij}.$$

Several of the estimated effects were trivially small, and so the final model was estimated excluding: $(\gamma_{01},\ldots, \gamma_{04},\gamma_{08},\gamma_{15},\gamma_{16},\gamma_{18},\gamma_{19},\gamma_{010},\gamma_{23},\gamma_{24},\gamma_{29},\gamma_{31},\gamma_{32},\gamma_{35},\ldots,\gamma_{310})$. The results for the reduced model are reported in Table 3 and discussed as follows.

Average School Achievement on Overall Provincial Examinations. The average years of parents' education were positively related to school mean achievement $(\hat{\gamma}_{02} = 1.22,\ t = 6.58)$, and so was the average number of examinations written in each school $(\hat{\gamma}_{03} = 1.34,\ t = 3.42)$. A modest positive relationship was indicated as well for school graduation rate $(\hat{\gamma}_{01} = 0.29,\ t = 6.44)$. Mean achievement was slightly lower in schools with high proportion of Aboriginal $(\hat{\gamma}_{04} = -0.05,\ t = -3.66)$, or Male students $(\hat{\gamma}_{05} = -0.06,\ t = -3.24)$.

Overall School Mark Differentiation. When the average overall provincial examination score was adjusted for the average school mark, there was a moderate positive relationship. The results suggest that there was considerable variability, on average, in the differentiating effects of school mark within schools across the two sectors. For some schools, this implies that the relationship between school marks and exam marks for

different types of students within a school differs significantly depending on the sector of the school attended ($\hat{\gamma}_{11} = .09,\ t = 3.38$).

The relationship between students on their marks and examination marks in independent schools was slightly stronger than for public schools. Upon closer analysis, it appears that students with low school marks in public schools, generally tended to have their scores adjusted slightly upwards on the overall provincial exam more than students with low school marks in independent schools. The differentiating effect of school mark was also positively related to the number of examinations written ($\hat{\gamma}_{15} = .17,\ t = 10.10$). The differentiating effect of school mark also varied across of the region vectors. This suggests that the magnitude of the impact school mark had within schools depended on the regional location of the schools compared. From the data, it appears that the school mark had a stronger differentiating effect in region vector R3 compared with the Lower Mainland than in the other vectors. In this vector a school was coded 1 if it was located in the Interior and Northern BC Region and is compared with the Lower Mainland that was coded 0 in all vectors.

ESL Status. There was a negative relationship between ESL status and the intercept ($\hat{\gamma}_{20} = -6.86,\ t = -13.54$). This implies that in the average school the ESL students typically score quite a bit lower; about 6.86 points lower than the non-ESL students in the average school. The relationship between ESL status and achievement across school sector was also negative ($\hat{\gamma}_{21} = -5.74,\ t = -2.7$), which implies that in the average independent school, the negative ESL achievement gap between students was much

Table 3

Results from the Intercepts- and Slopes-As-Outcomes for Example A

Fixed Effects	Coefficient	SE	t-ratio	p-value
School Exam mean achievement, β_{0j}				
Intercept, γ_{00}	68.14	0.16	421.68	0.000
Graduation rate, γ_{01}	0.29	0.05	6.44	0.000
Average years of parents' education, γ_{02}	1.22	0.19	6.58	0.000
Average number of exams, γ_{03}	1.34	0.39	3.42	0.001
Proportion Aboriginal, γ_{04}	-0.02	0.01	-3.66	0.001
Proportion Male, γ_{05}	-0.06	0.02	-3.24	0.002
School mark differentiation, β_{1j}				
Intercept, γ_{10}	0.64	0.01	86.79	0.000
Sector, γ_{11}	0.09	0.03	3.38	0.001
R1, γ_{12}	-0.05	0.02	-2.34	0.019
R2, γ_{13}	-0.09	0.02	-4.32	0.000
R3, γ_{14}	-0.14	0.02	-6.05	0.000
Average number of exams, γ_{15}	0.17	0.02	10.08	0.000
ESL slope, β_{2j}				
Intercept, γ_{20}	-6.86	0.51	-13.54	0.000
Sector, γ_{21}	-5.74	2.53	-2.27	0.023
R1, γ_{22}	-3.30	0.96	-3.46	0.001
Graduation rate, γ_{23}	0.26	0.09	2.87	0.005
Average years of parents' education, γ_{24}	-1.22	0.35	-3.49	0.001
Average number of exams, γ_{25}	3.65	0.72	5.05	0.000
Proportion ESL, γ_{26}	0.17	0.04	4.38	0.000
Proportion Male, γ_{27}	-0.15	0.07	-2.08	0.037
Gender slope, β_{3j}				
Intercept, γ_{30}	1.53	0.09	17.93	0.000
R2, γ_{31}	0.53	0.21	2.55	0.011
R3, γ_{32}	0.73	0.26	2.78	0.006
Random Effects	Variance	df	$\chi 2$	p Value
School mean achievement, μ_{oj}	5.86	248	2,914.07	0.000
School mark differentiation, μ_{1j}	0.008	248	899.72	0.000
Level-1 effect, r_{ij}	70.62			

larger than it would be within the average public school. The negative affect of the ESL status relationship is more pronounced for schools within region vector R1 ($\hat{\gamma}_{22} = -3.30,\ t = -3.46$) compared with the Lower Mainland. Also, as the proportion of males in a school increases, there was moderate negative relationship for ESL students. The relationship was positive as the average number of examinations written increased, when the school graduation rate was high, as well as if the proportion of ESL students in a school was large.

Gender Status. The impact of gender on the average provincial examination score was positive for males, and differed for two of the three region vectors (R2 and R3) compared with the Lower Mainland. It was positively related to both, with a stronger relationship for R3 ($\hat{\gamma}_{32} = .73,\ t = 2.78$) and slightly weaker for R2 ($\hat{\gamma}_{31} = .53,\ t = 2.55$). Therefore, these findings suggest the differentiation effect of gender status, or the "gender achievement gap", tended to be slightly larger within schools identified in region vector R3 (the Interior and Northern BC), as well as for schools located in R2 (Fraser Valley and Southern BC) when compared with schools located in the Lower Mainland Region.

Additional Information for Example A

In the intercepts- and slopes-as-outcomes model, the level-1 variance estimate, $\hat{\sigma}^2$, was almost the same as the random coefficient regression model. This was expected because the level-one variables did not change in the two models. The unconditional variance of intercepts was $\hat{\tau}_{00} = 14.12$ in the random coefficient regression model. It is considered a conditional or residual variance $\hat{\tau}_{00} = 5.86$ in the intercepts- and slopes-as-

outcomes model. The proportion of variance explained in β_{0j} and β_{1j} can be determined by applying equation 13.

$$\text{Proportion of Variance Explained in } \beta_{qj} = \frac{\hat{\tau}_{qq}(unconditional) - \hat{\tau}_{qq}(conditional)}{\hat{\tau}_{qq}(unconditional)}. \quad (13)$$

This implies that about 59% of the parameter variation in school average provincial examination mark (β_{0j}) has been explained by the inclusion of these school level variables in this model. The slope variability (β_{1j}) between schools on average school mark was reduced by 71%. The reliability estimates in this example decreased somewhat, but still suggested that there was considerable power left for examining the hypotheses about the effects of school characteristics on average provincial exam achievement as the intercept estimates were still highly reliable (λ = .887). A larger reduction occurred in the reliability of how the school characteristics influence the overall school mark differentiation (λ = .642), which would be expected due to the large reduction in slope variability. In Table 3, the χ^2 test statistic for the intercept takes on a value of 2,914 and for the school mark slope 2,168, both with 253 degrees of freedom. The null hypothesis is highly implausible ($p < .001$), indicating that significant variation does still exist among schools on average provincial examinations, and in the differentiating effect of school marks on the achievement on provincial examinations.

Empirical Bayes Estimates for Example A

As mentioned at the end of Chapter Three, parameters and residuals are difficult to predict in many models and even more so in higher levels of a multilevel model. Empirical Bayes estimates are calculated on the assumption that individuals belong to a

definable population where they can be assigned a quantity that will locate them within the underlying probability distribution of this statistical population. These estimates provide the best prediction of the unknown level-one parameter or residual for a particular school, which utilizes not only data from that specific school but also combines the data from all other similar schools to estimate each element of level-one. A good estimator minimizes the expected distance between the unknown parameter or residual and the estimator, which is why they are considered to be shrunken estimates (Raudenbush & Bryk, 2002). In the current study, this model would hypothesize that all students within a particular school have an unknown effect added to their expected score as a result of attending that school. With increasing reliability of the level-one estimates more weight is given to the school characteristics, which is based on information from the entire sample of schools. In contrast, if the level-one estimates are unreliable then more weight will be given to the school characteristics. The caution is to have an appropriate model specified for level-two. If the level-two model is misspecified – fails to include key aspects of educational policy and practice – then the estimates of the association between the student information and outcomes will be biased. Most school effectiveness researchers strongly recommend using confidence intervals around the point estimates to provide a more accurate picture.

It is clear from Figure 3 that the uncertainty attached to individual school estimates, at least from this data set, is that ranking schools would be extremely difficult. The schools were plotted according to their adjusted means calculated from the intercepts- and slopes-as-outcomes model, with their 95% confidence intervals that are portrayed by the upper and lower tails in Figure 3. As is demonstrated in the figure, these confidence intervals

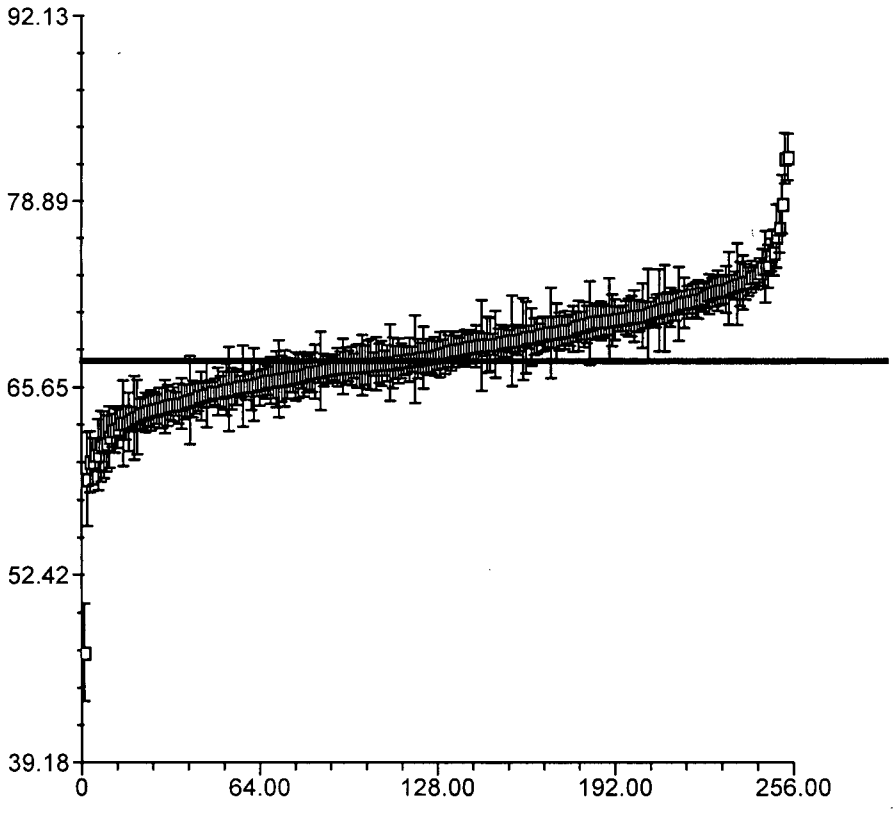

Figure 3. The Empirical Bayes Residual Point Estimates with 95% Confidence Intervals Plotted using HLM for Overall Examination Scores for Schools in Example A.

overlap the line indicating the overall average on the intercept (located on the Y axis) quite dramatically, which suggests that really only the group of the highest scoring schools and the lowest scoring schools (that do not have tails touching the line) could be used for comparisons. A further investigation of these data showed that the schools with the highest adjusted means were independent, same-sex female schools located in the Lower Mainland Region and the schools with lowest were mostly public, mixed gender, higher aboriginal proportion, and most were located in the Interior and Northern BC and Fraser Valley and Southern BC Regions. So, it would appear that even the lowest scoring schools should not be compared to the highest scoring schools because these are different types of schools.

Results for Example B – English Exam mark as Criterion

One of the main criticisms of the Fraser Institute report card was that it used the overall provincial examination results rather than a single subject area. Critics of the report card argued that schools might differ in their rank on the report card if different curriculum areas were reported. In order to test this assumption, the current study has selected the provincial examination curriculum area of English 12. The rationale for selecting the English 12 subject area was that it had the highest provincial participation rate of all the provincially examinable courses in this particular data set. Also, as mentioned in Chapter 3, the BC Ministry of Education claims that scores on this assessment demonstrate knowledge in one of the key provincial curriculum areas – literacy. For Example B, any schools with less than 10 students were excluded from the analyses, which resulted in three schools from Example A that were excluded in Example B. The same steps were followed for building the Example B intercepts- and slopes-as-

outcomes model, starting with an investigation of the descriptive statistics in Table 4, and then plotting the English examination marks and English school marks of students against a P-P plot to test the normality assumption shown in Figures 4 and 5.

One-Way ANOVA with Random Effects Model

As with Example A, the criterion variable for a one-way ANOVA with random effects model was fitted to the data for Example B, where the English 12 provincial examination mark was the criterion, and the residuals from each level are considered random. The average school mean, γ_{00}, was estimated as 70.34. The pooled within-school or level-one variance, $\hat{\sigma}^2$, was smaller than in Example A estimated at 131.92, and the variance among the 251 school means, $\hat{\tau}_{00}$, was also smaller at 10.47. The proportion of variance (or intra-class correlation) between schools can be determined using Equation 7. Therefore, about 7.4% of variance in the outcome can be attributed to differences between the schools and the remaining 92.6% to differences between students within schools. It appears again, that most of the variance in the outcome is explained by within-school variability rather than by attending a particular school. The 95% plausible values were estimated using Equation 8, and again had a wide range of about 12.7 points on average English exam achievement between schools (63.99, 76.69). The reliability of school sample means for this example was $\lambda = .867$, which implies for the schools in this sample there was considerable variability between schools on the outcome.

Random-Coefficient Regression Model

The same model outlined for Example A was applied to data for Example B in terms of the random-coefficient regression model. The results from this analysis are reported in Table 5.

Table 4

Descriptive Statistics for Example B with English Provincial Exam Score as Criterion

Variable Name	Mean	SD
Level-1 (35,887)		
Provincial English exam score	70.17	11.88
English School mark	72.66	12.92
ESL (0=Non-ESL, 1=ESL)	0.02	0.14
Gender (0=Female, 1=Male)	0.47	0.50
Level-2 (251)		
Average English provincial exam score	70.34	3.58
Sector (0=public, 1=independent)	0.15	0.36
R1 (1=Vancouver Island and the Coast)	0.20	0.40
R2 (1=Fraser Valley and Southern BC)	0.23	0.42
R3 (1=Interior and Northern BC)	0.18	0.39
School Graduation rate	94.84	3.60
Average years of parents' education	14.44	1.12
Average number of exams written	4.07	0.39
Proportion Aboriginal	4.67	9.62
Proportion ESL	1.42	3.18
Proportion Male	46.06	8.39

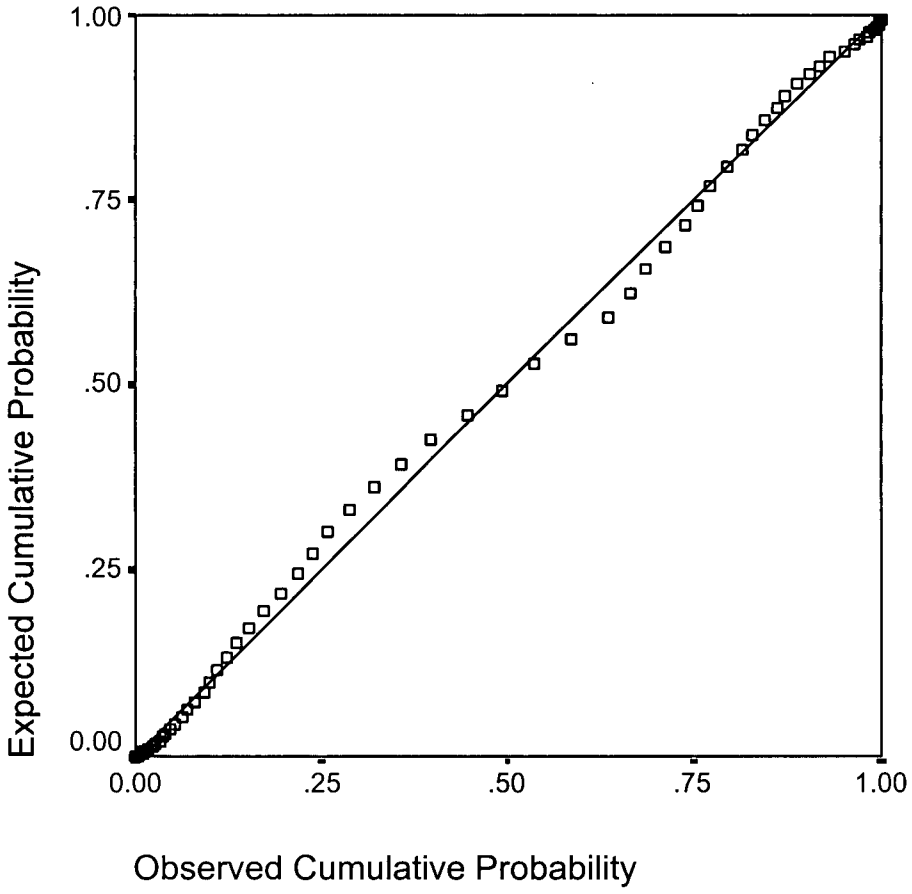

Figure 4. P-P Plot in SPSS to Determine Whether the Distribution of Student English Exam Mark is Normal in Example B with English 12 Examination Mark as the Criterion Variable.

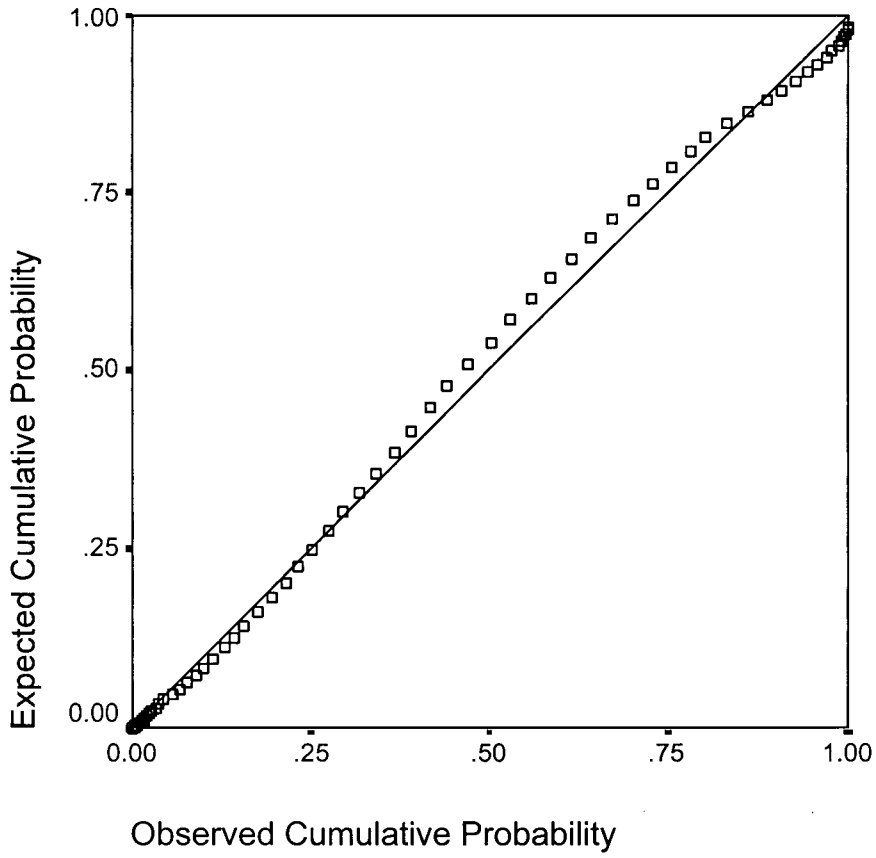

Figure 5. P-P Plot in SPSS to Determine Whether the Distribution of Student English School Mark is Normal in Example B.

The proportion of variance in level-one explained by this model was 38.2%. In this model, the χ^2 test statistic for the intercept was 4,468.91 and for the school mark slope was 1,18.50, both with 250 degrees of freedom. Therefore, the null hypothesis is highly implausible ($p < .001$), indicating again that significant variation does exist among schools on the achievement of English provincial examinations and on the differentiating impact of English school marks on the outcome. The plausible values were also estimated and as expected there was a wide spread among scores for school means (63.77, 76.91) as well as for school mark differentiation (.27, .81). There was only a small to moderate correlation among the school effects ($\hat{\rho} = 0.208$).

Table 5

Results from the Random Coefficient Regression for Example B

Fixed Effects	Coefficient	SE	t-ratio	p-value
Intercept, γ_{00}	70.34	0.22	318.70	0.000
Schlmark, γ_{01}	0.54	0.01	57.66	0.000
ESL, γ_{02}	-11.26	0.62	-18.11	0.000
Gender, γ_{03}	-.87	0.11	-8.08	0.000
Random Effects	**Variance Component**	**df**	**χ^2**	**p value**
School mean, μ_{oj}	11.23	250	4,468.91	0.000
Schlmark Slope, μ_{1j}	0.02	250	1,218.50	0.000
Level-1 effect, r_{ij}	81.51			
Reliability of OLS Regression-Coefficient Estimates				
School mean achievement	0.91			
School mark differentiation	0.71			

This was much lower when compared with Example A ($\hat{\rho} = 0.565$), most likely due to having more information provided on a single subject compared with an overall examination mark. The reliabilities for examining the hypotheses about the effects of school characteristics on average English exam achievement was high ($\lambda = .914$) and for studying how school characteristics influence the academic differentiation of English school mark ($\lambda = .707$).

Intercepts- and Slopes-as-Outcomes Model

As with Example A, the random coefficient regression model for Example B also indicated that each of the level-one predictors had a statistically significant relationship with the English provincial examination scores. The same intercepts- and slopes-as-outcomes full model was applied to this second data set, and the best-fitted model did not include the variables ($\gamma_{06},\ldots,\gamma_{010},\gamma_{16},\ldots,\gamma_{110},\gamma_{22},\ldots,\gamma_{210},\gamma_{32},\ldots,\gamma_{310}$), that were not statistically significant. The results for the reduced model are reported in Table 6.

Average English Provincial Exam Achievement. The school sector was positively related to average English provincial exam achievement, which differed from the results found in Example A where the criterion was not statistically significantly related to sector. In Example B, students attending an independent school will generally score slightly higher on their English provincial examination than students attending a public school ($\hat{\gamma}_{01} = 1.42,\ t = 2.42$). The relationship of the outcome in this model was similar to that in Example A with respect to school graduation rate and average years of parents' education. As might be expected on an English examination, as the proportion of ESL students in a school increased, there was a negative relationship with the outcome

($\hat{\gamma}_{04} = -.22,\ ,t = -3.84$). A small negative relationship also existed as the proportion of males increased in a school.

Overall English School Mark Differentiation. The school mark differentiation effect varied for the two sectors, even more so than it did in Example A ($\hat{\gamma}_{11} = .18,\ t = 6.30$). This implies that students attending an independent school will have a stronger differentiating influence between their English school mark and their English provincial examination mark than students attending the typical public school in this sample. Again, upon closer examination of these data it seems that students with low English school marks attending a public school, on average will score higher on the English provincial examination than students in independent schools. As with Example A, the English school mark differentiation varied across regions in a negative relationship, again with a stronger impact in region vector R3 (Interior and Northern BC). This implies that there is less of a school mark differentiating effect within schools located in regions outside the Lower Mainland, especially for schools within the Interior and Northern BC. There was a small positive relationship with average years of parents' education ($\hat{\gamma}_{15} = 0.02,\ t = 2.97$), which did not exist for Example A.

ESL Status. The only impact of ESL status on achievement was a positive relationship as the proportion of ESL students in a school increased. The contextual effect of ESL status in a school with the average proportion of ESL students would decrease the negative relationship between the ESL gap and English exam achievement from –13.60 to –11.87. All other variables for ESL status were not statistically significant.

Table 6

Results from the Intercepts- and Slopes-as-Outcomes for Example B

Fixed Effects	Coefficient	SE	t-ratio	p-value
English mean achievement, β_{0j}				
Intercept, γ_{00}	70.40	0.18	391.90	0.000
Sector, γ_{01}	1.42	0.58	2.42	0.016
Graduation rate, γ_{02}	0.23	0.06	4.07	0.000
Average years of parents' education, γ_{03}	1.04	0.17	6.08	0.000
Proportion ESL, γ_{04}	-0.22	0.06	-3.84	0.000
Proportion Male, γ_{05}	-0.10	0.02	-4.24	0.000
School mark differentiation, β_{1j}				
Intercept, γ_{10}	0.54	0.01	68.27	0.000
Sector, γ_{11}	0.18	0.03	6.30	0.000
R1, γ_{12}	-0.09	0.02	-4.28	0.000
R2, γ_{13}	-0.08	0.02	-3.68	0.000
R3, γ_{14}	-0.12	0.02	-4.98	0.000
Average years of parents' education, γ_{15}	0.02	0.01	2.98	0.004
ESL slope, β_{2j}				
Intercept, γ_{20}	-13.60	0.52	-26.18	0.000
Proportion ESL, γ_{21}	0.37	0.06	6.24	0.000
Gender slope, β_{3j}				
Intercept, γ_{30}	-0.80	0.10	-7.92	0.000
Graduation rate, γ_{31}	0.08	0.03	2.67	0.008
Random Effects	Variance Component	df	$\chi 2$	p Value
School mean achievement, μ_{oj}	7.06	245	2,792.09	0.000
School mark differentiation, μ_{1j}	0.01	245	845.50	0.000
Level-1 effect, r_{ij}	81.42			

Gender Status. The only influence gender status had on the average English provincial examination results was a positive relationship with graduation rate ($\hat{\gamma}_{31} = 0.08,\ t = 2.67$). As the graduation rate of a school increased, the gender slope was positively influenced.

Additional Information for Example B

The unconditional variance of intercepts was $\hat{\tau}_{00} = 11.23$ in the random-coefficient regression model. The conditional variance in the intercepts- and slopes-as-outcomes model was $\hat{\tau}_{00} = 7.06$. The proportion of variance explained in β_{0j} and β_{1j} can be calculated using Equation 11. Therefore, about 37% of the parameter variation in school average English provincial examination was explained. The slope variability between schools on average English school mark was reduced by 40%. The reliability estimate for the intercept decreased slightly ($\lambda = .875$) and so did the reliability for the slope ($\lambda = .610$). The χ^2 test statistic for the intercept was 2,792 and for the slope was 845, both with 245 degrees of freedom. This indicates that we still are unable to accept the null hypothesis because significant variation ($p < .001$) still exists among schools in this data set on the English provincial examination scores that has not yet been explained.

Empirical Bayes Estimates for Example B

Figure 6 shows again that the confidence intervals overlap for most schools in the sample so that detailed rank orderings are statistically invalid for a majority of the schools in this sample. The highest scoring schools on the English Examination were very similar in type to those scoring high on the overall provincial examination mark.

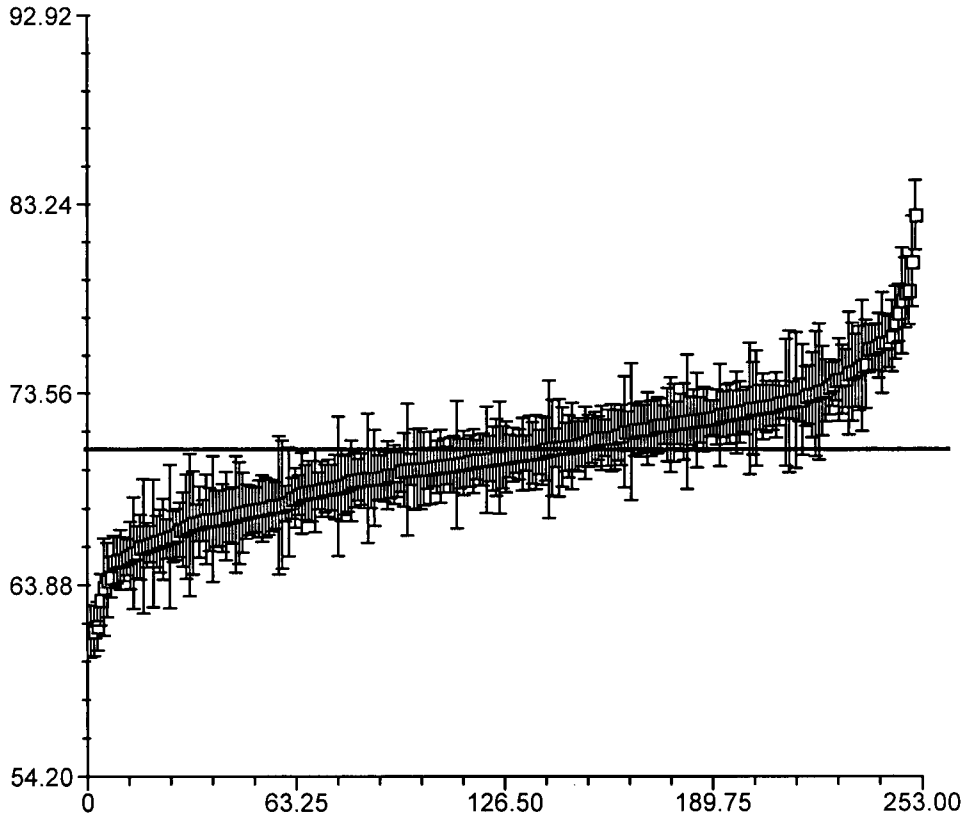

Figure 6. The Empirical Bayes Point Estimates with 95% Confidence Intervals Plotted using HLM for English Examination Scores for Schools in Example B.

This provides visual evidence for the findings in Table 6 that sector does have a significant impact on the intercept when the criterion variable is a single subject area rather than an overall examination score.

Summary of Main Findings

The principal aim of these analyses has been to demonstrate how differences between schools in examination results vary by the student-level and school-level characteristics. In the current study it was assumed that the overall provincial examination score (Example A) and the English provincial examination score (Example B) achieved by a student, could be in part attributed to the school and described as a school-level characteristic. Schools in these analyses were considered separate entities, and they were only connected in the second level of the model. For this purpose, it made sense theoretically to group-mean centre all the level-one characteristics. By group-mean centering school mark, ESL status, and gender, in level one, then achievement on the outcome can be seen as a relative effect that is partly determined by school factors. This was discussed in Chapter 3 as the frog-pond effect, which implies that results on the outcome will depend on whether a student is a small or large frog in the pond, as well as the size of the pond, type and location. The proportion of ESL students and the proportion of males within a school were also included in level two of the model and were grand-mean centred. This was done to achieve additional information about whether the influence of these second level variables - proportion of ESL and male students - could explain the left over variation between schools that remained unexplained by the level-one variables.

The intercepts- and slopes-as-outcomes analyses for both Example A and Example B statistically demonstrated why some schools have higher means than others on the outcome, and why in some schools the association between school mark and overall exam mark was stronger than in others. Furthermore, by introducing two different outcome models (Example A compared with Example B), it was easy to see that school results varied according to the selection of the criterion variable. This was made clear from the results that demonstrated the relationship between sector and the intercept was not found to be statistically significant for Example A, but was for Example B. Two other variables whose relationships with the intercept differed between the two models were the proportion of Aboriginal students, and the proportion of ESL students. There was a statistically significant relationship with school mean achievement and proportion of Aboriginal students on the overall provincial examination in Example A, whereas in Example B the proportion of Aboriginal students did not have a statistically significant relationship with the overall examination mark. The reverse was true for the relationship to the intercept for the proportion of ESL students across the two models. This might not be too unexpected, as the effects of proportion of ESL students in a school may have been averaged out for the overall provincial examination, but were not for the English examination. It would be anticipated that if an additional outcome variable on a single subject were to be included, such as Principles of Mathematics 12, the results would differ yet again. In line with what could be expected, the average years of parents' education and school graduation rate were positively related to the outcome for both models. The proportion of males was negatively related to the intercept in both examples, with a stronger negative relationship in Example B.

The findings indicated that the impact of school mark differentiation varied within schools across sector and region for Example A and B. This implies that the relationship of school mark and provincial examination mark varied in its strength between students within a school depending on which school type attended (independent compared with public) as well as its regional location in comparison to the Lower Mainland. The student-level characteristic of ESL status was statistically significantly related to more school-level variables in Example A than in Example B. The strength of the ESL achievement gap within schools for the average English 12 examination mark was more substantial than it was for the overall provincial examination mark. The ESL achievement gap was larger for ESL students attending the average independent school than it was for ESL students attending the average public school. In comparison to the Lower Mainland, positive regional differences occurred for the gender achievement gap in Example A for schools located on Vancouver Island and the Coast (vector R2), and within the Interior and Northern BC (vector R3). This indicates that male students perform, on average, better compared with their female schoolmates in these regions compared with students attending Lower Mainland schools. There were no statistically significant findings across regions for Example B, but graduation rate had a positive association with the gender achievement gap in this example. Overall, the results of these two examples suggest that schools in these datasets varied along many dimensions and that the majority of schools cannot be statistically distinguished from each other. Those that appear to be able to be distinguished (those ranking really low and really high in Figures 3 and 6 do not look like they are the same types of schools and so further investigation would be required prior to comparison.

Chapter Five

Discussion and Conclusions

Principal Findings of the Study

A reasonable amount of the variance between schools and students on the outcomes was explained for both Example A and Example B by building two intercepts- and slopes-as-outcomes models. However, even with the inclusion of student characteristics and school contextual characteristics, significant differences between outcomes remain unexplained. The results from these two intercepts- and slopes-as-outcomes models demonstrated how the relationship between the student-level characteristics and the school contextual variables differ quite substantially depending on the criterion selected. Furthermore, the findings uncovered that the strength of the differentiating impact of school mark, as well as the magnitude of the ESL and gender achievement gaps, depended on the sector and regional location of schools.

Example A - Overall Average Examination Mark as the Criterion

When the overall average examination mark was included as the criterion variable, the statistically significant relationships between the school-level explanatory variables and the achievement on examination were not surprising. However, it was hypothesized that sector and regions would have statistically significant relationships with the outcome, which did not emerge with these data in Example A. When investigating the differentiating effects of school mark within this model, statistically significant relationships for sector and regions with the outcome did materialize. This suggests that the influence of school mark in predicting a student's score on the final examination will be determined by individual-level characteristics of the student, what school sector (whether independent or public) that student attends, as well

as the region in which the school is located. Uncovering this level of information from the data is the unique benefit of applying a multilevel approach. With reference to the previously mentioned frog-pond effect, the influence of school mark will depend on the personal characteristics of the frog, and on the type, location and size of the pond. Another finding that was revealed in Example A is that public schools, on average, have a smaller ESL achievement gap than independent schools on achievement for the overall provincial examination. Furthermore, the Lower Mainland schools have a smaller ESL achievement gap for students, on average, than schools located on Vancouver Island and the Coast. Both results could possibly be explained by school or district policy. In those data, there were typically more ESL students enrolled in public schools than in independent schools. The same was true for schools located in the Lower Mainland compared with other regions. As the proportion of ESL students increases in a school or district, it could be expected that special programming would be implemented in order to support ESL students academically. Thus, where the proportion of ESL students is greater, it would be expected that these special programs would reduce the ESL achievement gap in comparison to other schools, which is what the data seemed to imply. The results also suggested that for Example A there was a positive influence of being male on the outcome. Among these groups of schools included in the sample, males typically scored higher than their female schoolmates, and the gap in favour of males was wider in Fraser Valley/Southern Interior region, and to a much greater extent in the Interior/Northern region of BC. The results of the analysis from Example A were quite different when the criterion variable was changed to English 12 examination score.

Example B - Average English Examination Mark as the Criterion

As was originally hypothesized, school sector did emerge as a statistically significant

influence on the outcome when the average English 12 examination mark was adopted as the criterion. Based on that result, one may infer that the average English examination mark for independent schools in this sample were typically higher than expected for the overall provincial examination when compared with the average public school. Therefore, by selecting a single subject area as the criterion, underlying relationships emerged that did not when the overall provincial examination mark was used in the analysis. School sector also had a stronger school mark influence in Example B in comparison with Example A. In Example A, the findings suggested that for students attending an independent school, their school marks were typically more closely related to their final examination mark as compared with students attending a public school. A similar relationship was found in Example B, and it was even stronger. For both models, the influence of school mark tended to be stronger for schools in the Lower Mainland compared with all other regions. It might be concluded that public schools are inclined to conduct more broad-based assessments of students or tend to offer school-based curriculums that are more extensive than the content covered on the provincial examination. However, the reasons for the differences between sectors would require further investigation. As might be expected, the width of the ESL achievement gap within schools on the average English 12 examination was more substantial than it was for the overall provincial examination. This large negative achievement gap for ESL students compared to non-ESL students appeared regardless of the context of the school; except that it was decreased as the proportion of ESL students within a school increased. A negative gender achievement gap for males also appeared regardless of most of the school level variables included in the model. This was a reverse of the positive gender gap that emerged in Example A. However, for Example B it was made smaller as the graduation rate of the

school increased.

Discussion: Implications of Comparing BC Schools

Overall, the results of the two models suggest that schools in these datasets varied along many dimensions and that the majority of schools in these samples cannot be statistically distinguished from each other based only on examination results. Those that appear to be able to be distinguished (i.e., those ranking really low and really high in Figures 3 and 6 in Chapter 3) do not seem to be the same types of schools. Therefore, further investigation would be required prior to comparison. For some schools, there are substantial differences in the average achievement depending on the individual characteristics of the student and characteristics of the school. A male student attending a public school in the Lower Mainland region will have quite a different achievement experience compared with a male student attending a public school in the Interior and Northern BC region. The experience would change yet again if a male student attending a public school were compared with a male student attending an independent school. Therefore, a single effectiveness measure that only investigates the average school examination mark across school sectors and regions may mask important within school and between school differences that occur for different types of students within different types of schools.

Accountability is of paramount importance to the current BC provincial government for informing system-wide initiatives and policy planning, as well as informing judgments about individual public schools. Therefore, it is imperative that the definition of effectiveness within this accountability framework be clearly stated and understood. Additional studies, such as the current one, should be conducted to provide valuable information to the decision-makers in the BC Ministry of Education. The authors of the report card published by Fraser

Institute, Cowley and Easton, should be recognized for their initial steps in their efforts to understand the relationships between schools and achievement on the provincial examinations. They have acknowledged a demand for this type of information and have adopted it with more fervor than any other organization in BC. Authors of the report card suggested that school comparisons are at the heart of the improvement process and that there is great benefit in identifying schools that are particularly effective. However, based on the findings of the current study, the relationships are much more complicated than they appear in the report card. First, report card results only investigate the aggregated between-school differences. Second, report card results missed some important within-school relationships that affect these between-school differences.

With reference to Chapter One, the principal assumption underlying the theoretical framework of the current multilevel approach was that schools are not separate, self-enclosed, and self-referential institutions. Students are nested in schools, which are nested in communities, and all are interconnected. Under that premise, the present study aimed to investigate three research questions. First, how much of the variability in school performance on overall Grade 12 provincial examinations, as well as school performance on one provincially examinable subject area – English 12 – could be attributable to differences between schools and how much to differences between students within schools? Second, to what extent does the school attended influence the students' academic attainments? Third, are there factors at the student and school levels that account for the variability at either level (contextual and cross-level relationships)? The findings in the present study highlight how these samples of secondary schools in BC differ quite substantially on examination achievement, and how including student-level information and information about school

context, identifies the complicated relationships occurring within and between schools on these assessments. These relationships are hidden when using ordinary linear multiple regression models that incorporate only aggregated school-level data. The results from the current study, for both examples, demonstrated an enormous benefit of using multilevel models when investigating differences between schools, as well as the limitations to using report cards based on a single numerical score for comparing schools in BC. As the findings suggest, schools in BC should not be compared to each other without taking into consideration their many differences. Figures 3 and 6 illustrate that when schools were ranked according to their empirical Bayes estimates with 95% confidence intervals, the majority of schools could not be statistically distinguished from each other. In other words, the portrait of school effectiveness, based on the English provincial and overall provincial examination for secondary schools in BC, must include the many relationships that are multi-dimensional in order to portray a more accurate picture of what is expected of a school performing effectively.

Limitations of the Study

A few important relationships among Grade 12 students and schools involved in the BC provincial examination program in the academic year 2002/03 were revealed through the present study. However, as mentioned previously, there were still unexplained differences between and within schools for both models. Resources and available data limited the scope of this current investigation. For those reasons, five main limitations are discussed that could provide additional information about the achievement differences between schools. These include: 1) individual-level data on the socio-economic status of students, 2) value-added or baseline information, 3) classroom and teacher-level variables, 4) non-academic subject

variables as the criterion or predictor, and 5) additional years' worth of data for longitudinal investigation.

Individual socio-economic status indicators (SES) were not included because the costs of obtaining those data from Statistics Canada were beyond the reach of the investigator. Therefore, only one school-level variable that was collected from the published Fraser Institute's 2004 Edition of the report card was included in both models and used as a proxy for school-level SES. Of course, it would be more desirable to have the student-level data collected from Statistics Canada (by way of forward sortation areas matched to student-level postal codes) and then through multiple regression analyses determine the most important SES variables to include in the final multilevel model.

Another limitation of the present study was that there were no value-added data included in these models, which limits the amount of information obtained for investigating the differential effectiveness of a particular school. Without the addition of intake variables there were no controls for baseline or initial capacities at the student level within schools. This would mean that there could be unidentified statistically significant occurrences that were missed because they are contained within the information that existed prior to the extraction of these data. Intake and developmental factors may interact with external factors and with the characteristics of schools. According to the research reviewed in Chapter Two, it would be expected that with the addition of value-added or baseline information, more unexplained variance between schools could be explained.

The present study was also unable to investigate classroom-level and teacher-level effects as the intermediary level between students and schools. In essence, this level was purposively ignored for the current study. Findings from school effectiveness research

suggest that when classrooms are included as a level between the student and the school, the between-classroom variation in achievement is quite large, sometimes even more than what was found between schools. In some situations, the school variations were reduced to a very small amount when including the additional level of classroom or school (Waxman & Huang, 1997). The classroom and teacher levels should be identified because learning takes place in the classroom as well as within schools.

The criterion variables included in this study were all based on the provincially examinable program for Grade 12 students enrolled in BC secondary schools. However, information on how effective schools are should not be reduced to only academic curriculum. There are schools that perform quite effectively on non-provincially examinable programs, such as in music, theatre, industrial trades, or specialized areas of physical education. Furthermore, by only including paper-and-pencil type assessments the entire range of programming offered in BC schools is not captured.

The models developed in this investigation are inherently constrained to answer the research question, "How does attending a particular school influence academic achievement on the overall provincial examination and the English 12 provincial examination?". The development of different types of models, compared to the ones employed in this study, would be required to answer questions about how schools improve over time. In addition, these types of studies would require at least three years of longitudinal data, and the current study only included one year's worth of data. A study that investigated improvement over time should also contain value-added or baseline information.

Recommendations for Further Research

As was highlighted above, research investigating how schools differ from each other on academic achievement can expand much beyond the scope of the current study. Suggestions for further research would include first adding student-level socio-economic characteristics and some value-added information to provide a baseline from which a school can be deemed internally effective or ineffective. This would also help in determining the potential for schools to perform differently for students with a variety of background characteristics. Also, it would be likely that by incorporating classroom-level and teacher-level effects into a three-level multilevel model, more of the within-school variability could be explained. According to the findings from this current study, this would be of particular interest in order to investigate why there appeared to be differentiating influences of school mark within schools depending on the sector and regional location of schools.

A model that incorporated value-added information, student-level SES, non-academic subjects, and a third level (classroom or teacher) into the framework used in the present study, as well as at least three years' worth of data, would provide a better gauge of school effectiveness and provide more valuable information to help guide school-based initiatives and system-wide policy planning.

Bibliography

Aitkin, M. & Longford, N. (1986). Statistical Modelling Issues in School Effectiveness Studies. Journal of the Royal Statistical Society Series A (Statistics in Society), 149, 1, 1-43.

Aguinis, H. (July 2002). Estimation of Interaction Effects in Organization Studies. Organizational Research Methods, 5, 3, 207-211.

Austin, G. R. (October 1979). Exemplary Schools and the Search for Effectiveness. Educational Leadership, 37, 10-14.

Ballou, D., Sanders, W., & Wright, P. (Spring 2004). Controlling for Student Background in Value-Added Assessment of Teachers. Journal of Educational and Behavioral Statistics, 29, 1, 37-65.

BC Ministry of Education. (2003a). Enhancing Learning: Report of the Student Achievement Task Force. Submitted to Honourable Christy Clark, Minister of Education by F. Brownlie, S. Ladyman, J. MacRae, F. Renihan, G. Sumanik, and R. Wickstrom (Chair).

BC Ministry of Education. (2003b). Enhancing Rural Learning: Report of the Task Force on Rural Education. Submitted to Honourable Christy Clark, Minister of Education by H. Clarke, J. Imrich (Chair), E. Surgenor, and N. Wells.

Bickel, R. & Howley, C. (May 2000). The Influence of Scale on School Performance: A Multi-Level Extension of the Matthew Principle. Education Policy Analysis Archives, 8, 22, 3-33.

Blatchford, P., Moriarty, V., Edmonds, S., Martin, C. (Spring 2002). Relationships

Between Class Size and Teaching: A Multimethod Analysis of English Infant Schools. American Educational Research Journal, 39, 1, 101-132.

Bock, D. (1989). Multilevel Analysis of Educational Data. New York, NY: Academic Press, Inc.

Burke, J., Minassians, H., & Yang, P. (Fall 2002). State Performance Reporting Indicators: What Do They Indicate? Planning for Higher Education, 15-29.

Carver, Ronald P. (1975). The Coleman Report: Using Inappropriately Designed Achievement Tests, American Educational Research Journal, 12, 1, 77-86.

Chester, M. (Summer 2003). Multiple Measures and High Stakes Decisions: A Framework for Combining Measures. Educational Measurement: Issues and Practice, 22, 2, 32-41.

Cistone, P. & Bashford, J. (Summer 2002). Toward a Meaningful Institutional Effectiveness Plan: Learning from Accreditation. Planning for Higher Education, 30, 4, 15-23.

Coe, R. & Fitz-Gibbon, C.T. (December 1998). School Effectiveness Research: Criticisms and Recommendations. Oxford Review of Education, 24, 4, 421-438.

Coleman, J.S., Campbell, E.Q., et al. (1966). Equality of Educational Opportunity. Washington, DC: U.S. Government Printing Office.

Cowley, P. & Easton, S. (March 2004). Report Card on British Columbia's Secondary Schools: 2004 Edition. The Fraser Institute, Vancouver, BC, Canada.

Cronbach, L. & Linn, R. (June 1997). Generalizability Analysis for Performance Assessments of Student Achievement or School Effectiveness. Educational and Psychological Measurement, 57, 3, 373-399.

Davison, M., Kwak, N., Seok Seo, Y., and Choi, J. (July 2002). Using Hierarchical Linear Models to Examine Moderator Effects: Person-by-Organization Interactions. Organizational Research Methods, 5, 3, 231-254.

Dowsett Johnston, A. (November 2001). Choosing the Right University: An Insider's Guide. Maclean's Magazine, 114, 47, 22-64.

Edmonds, R. (October 1979). Programs of School Improvement: An Overview. Educational Leadership, 37, 20-24.

Freeman, J., Teddlie, C., & Kennedy, E. (1998). A Longitudinal View of Change in School Effectiveness Status. Retrieved January 15, 2001, from http://www.wcer.wisc.edu/sipsig/freeman.html

Gibson, A., & Asthana, S. (June 1998). School Performance, School Effectiveness and the 1997 White Paper. Oxford Review of Education, 24, 2, 195-210.

Goldstein, H. (1997). Methods in School Effectiveness Research. School Effectiveness and School Improvement, 8, 369-395.

Goldstein, H. (2001). Using Pupil Performance Data for Judging Schools and Teachers: Scope and Limitations. British Educational Research Journal, 27, 4, 433-442.

Goldstein, H., Huiqi, P., Rath, T., & Hill, N. The Use of Value Added Information in Judging School Performance. Retrieved August 17, 2004 from http://www.ioe.ac.uk/hgpersonal/Using-value-added-information.pdf

Goldstein, H. & Rasbash, J. (1993). A Multilevel Analysis of School Examination Results. Oxford Review of Education, 19, 4, 425-433.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H. Nuttall, D. & Thomas, S.

(June 1992). Multilevel Models for Comparing Schools. Multilevel Modelling Newsletter, 4, 2, 5-6.

Goldstein, H. and Speigelhalter, D.J. (1996). League Tables and their Limitations: Statistical Issues in Comparisons of Institutional Performance. Journal of the Royal Statistical Society, Series A (Statistics in Society), 159, 3, 385-443.

Goldstein, H. & Thomas, S. (1996). Using Examination Results as Indicators of School and College Performance. Journal of the Royal Statistical Society Series A (Statistics in Society), 159, 1, 149-163.

Goldstein, H. & Woodhouse G. (2000). School Effectiveness Research and Educational Policy. Oxford Review of Education, 26, 3&4, 353-363.

Gray, J., Goldstein, H., and Thomas, S. (2001). Predicting the Future: The Role of Past Performance in Determining Trends in Institutional Effectiveness at A Level. British Educational Research Journal, 27, 4, 391-405.

Gray, J., Jesson, D. (June 1990). Estimating Differences in the Examination Performances of Secondary Schools in Six LEAS: A Multi-Level Approach to School Effectiveness. Oxford Review of Education, 16, 2, 137-158.

Gray, J., Jesson, D., Goldstein, H., Hedger, K., and Rasbash, J. (1995). A Multi-Level Analysis of School Improvement: Changes in Schools' Performance Over Time. School Effectiveness and School Improvement, 6, 2, 97-114.

Hargreaves, D. (2001). A Capital Theory of School Effectiveness and Improvement. British Educational Research Journal, 27, 4, 487-503.

Heck, R. & Thomas, S. (2000). An Introduction to Multilevel Modeling Techniques. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hill, P. & Rowe, K. (March, 1995). Use of Multi-Level Modelling in Procedures for Maximizing Between-School Comparability of Final Year School-Based Assessments. Multlevel Modelling Newsletter, 7, 1, 12-16.

Hofmann, D. & Gavin, M. (1998). Centering Decisions in Hierarchical Linear Models: Implications for Research in Organizations. Journal of Management, 24, 5, 623-641.

Hopkins, D., Reynolds, D., Gray, J. (1999). Moving on and Moving up: Confronting the Complexities of School Improvement in the Improving Schools Project. Educational Research and Evaluation, 5, 1, 22-40.

Jencks, Christopher S. & Brown, M. (August 1975). The Effects of High Schools on Their Students. Harvard Educational Review, 45, 3, 273-324.

Kezar, A. & Eckel, P. (July/August 2002). The Effect of Institutional Culture on Change Strategies in Higher Education: Universal Principles or Culturally Responsive Concepts? The Journal of Higher Education, 73, 4, 435-460.

Kreft, I.G.G. (December, 1995). The Effects of Centering in Multilevel Analysis: Is the Public School the Loser or the Winner? A New Analysis of an Old Question. Multilevel Modelling Newsletter, 7, 3, 5-8.

Kreft, I.G.G., de Leeuw, J., & Aiken, L. (1995). The Effect of Different Forms of Centering in Hierarchical Linear Models. Multivariate Behavioral Research, 30, 1, 1-21.

Lane, S. & Stone, C. (Spring 2002). Strategies for Examining the Consequences of Assessment and Accountability Programs. Educational Measurement: Issues and Practice, 21, 1, 23-30.

Linn, R. (December 2001). Reporting School Quality in Standards-Based Accountability Systems. National Council of Measurement in Education Newsletter, 9, 4, 4-5.

Linn, R., Baker, E., & Betebenner, D. (September 2002). Accountability Systems: Implications of Requirements of the No Child Left Behind Act of 2001. Educational Researcher, 21, 1, 23-30.

Lockwood, J. R., Louis, T. A., McCaffrey, D. F. (Fall 2002). Uncertainty in Rank Estimation: Implications for Value-Added Modeling Accountability Systems. Journal of Educational and Behavioral Statistics, 27, 3, 255-270.

McCaffrey, D., Lockwood, J.R., Koretz, D., Louis, T. A., Hamilton, L. (Spring 2004). Models for Value-Added Modeling of Teacher Effects. Journal of Educational and Behavioral Statistics, 29, 1, 67-101.

Mehrens, W. A. (Spring 1992). Using Performance Assessment for Accountability Purposes. Educational Measurement: Issues and Practices, 3-9.

Meng Thum, Y. & Bhattacharya, S. (Winter 2001). Detecting a Change in School Performance: A Baysian Analysis for a Multilevel Join Point Problem. Journal of Educational and Behavioral Statistics, 26, 4, 443-468.

Messick, S. (September 1995). Validity of Psychological Assessment: Validation of Inferences From Persons' Responses and Performances as Scientific Inquiry Into Score Meaning. American Psychologist, 50, 9, 741-749.

Michell, J. (1999). Measurement in Psychology: A Critical History of A Methodological Concept. Cambridge, UK: Cambridge University Press.

Mok, M. (June 1995). Sample Size Requirements for 2-Level Designs in Educational Research. Multilevel Modelling Newsletter, 7, 2, 11-15.

Morrison, H.G. & Cowan, P.C. (April 1996). The State Schools Book: A Critique of A League Table. British Educational Research Journal, 22, 2, 241-250.

Nash, R. (2001). Progress at School and School Effectiveness: Non-cognitive Dispositions and Within-class Markets. Journal of Education Policy, 16, 2, 89-102.

Opdenakker, Marie-Christine and Van Damme, Jan. (2000). Effects of Schools, Teaching Staff and Classes on Achievement and Well-Being in Secondary Education: Similarities and Differences Between School Outcomes, School Effectiveness and School Improvement, 11, 2, 165-196.

Opdenakker, Marie-Christine and Van Damme, Jan. (2001). Relationship between School Composition and Characteristics of School Process and their Effect on Mathematics Achievement, British Educational Research Journal, 27, 4, 407-432.

Orsak, Mendro, & Weerasinghe. (1998). Calculating Missing Student Data in Hierarchical Linear Modeling: Uses and Their Effects on School Rankings. Multiple Regression Viewpoints, 25, 3-12.

Osborne, J. (2000). Advantages of Hierarchical Linear Modeling. Practical Assessment, Research and Evaluation, 7, 1, 1-7.

Paterson, L., & Goldstein, H. (1991). New Statistical Methods for Analysing Social Structures: An Introduction to Multilevel Models, British Educational Research Journal, 17, 4, 387-393.

Pituch, K. (April 1999). Describing School Effects with Residual Terms: Modeling the Interaction Between School Practice and Student Background. Evaluation Review, 23, 2, 190-211.

Raudenbush, S. (Spring 2004). What Are Value-Added Models Estimating and What Does This Imply for Statistical Practice? Journal of Educational and Behavioral Statistics, 29, 1, 121-129.

Raudenbush, S. & Bryk, A. (2002). Hierarchical Linear Models: Applications and Data Analysis Methods Second Edition. Thousand Oaks, CA: Sage Publications, Inc.

Raudenbush, S. & Willms, J. (1995). The Estimation of School Effects. Journal of Education and Behavioral Statistics, 20, 307-335.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I., and Lewis, T. (July 2002). A User's Guide to *MLwiN.* Version 2.1d, Centre for Multilevel Modelling Institute of Education, University of London.

Reynolds, D. & Teddlie, C. (2001). Reflections on the Critics, and Beyond Them. School Effectiveness and School Improvement, 12, 1, 99-113.

Sammons, P. & Nuttall, D. (1993). Differential School Effectiveness: Results from a Reanalysis of the Inner London Education Authority's Junior School Project Data. British Educational Research Journal, 19, 4, 381-392.

Schafer, W. (Summer 2003). A State Perspective on Multiple Measures in School Accountability. Educational Measurement: Issues and Practice, 22, 2, 27-31.

Schagen, I.P. (1990). Analysis of the effects of school variables using multi-level models. Educational Studies, 16, 1, 61-73.

Schagen, I. & Hutchison, D. (October 2003). Adding Value in Educational Research -

The Marriage of Data and Analytical Power. British Educational Research Journal, 29, 5, 749-765.

Scheerens, J., Bosker, R., & Creemers, B. (2001). Time for Self-Criticism: on the Viability of School Effectiveness Research. School Effectiveness and School Improvement, 12, 1, 131-157.

Schmidt, W., McKnight, C., Houang, R., Wang, H., Wiley, D., Cogan, L., & Wolfe, R. (2001). Why Schools Matter: A Cross-National Comparison of Curriculum and Learning. San Francisco, CA: John Wiley & Sons, Inc.

Scott, R. & Walberg, H. (October 1979). Schools Alone are Insufficient: A Response to Edmonds. Educational Leadership, 37, 24-27.

Singer, J. (1998). Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models. Journal of Educational and Behavioral Statistics, 24, 4, 323-355.

Slee, R. & Weiner, G. (March 2001). Education Reform and Reconstructions as a Challenge to Research Genres: Reconsidering School Effectiveness research and Inclusive Schooling. School Effectiveness and School Improvement, 12, 1, 83-98.

Spencer, N.H. & Fielding, A. (2002). A Comparison Of Modelling Strategies For Value-Added Analyses Of Educational Data. Computational Statistics, 17, 1, 103-116. ISSN 0943-4062.

Tate, R. (2004). Interpreting Hierarchical Linear and Hierarchical Generalized Linear Models With Slopes as Outcomes. The Journal of Experimental Education, 73, 1, 71-95.

Teddlie, C. & Reynolds, D. (2001). Countering the Critics: Responses to Recent Criticisms of School Effectiveness Research. School Effectiveness and School Improvement, 12, 1, 41-82.

Tekwe, C., Carter, R., Ma, Chang-Ma, Algina, J., Lucas, M., Roth, J., Ariet, M., Fisher, T., and Resnick, M. (Spring 2004). An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance. Journal of Educational and Behavioral Statistics, 29, 1, 11-36.

Thomas, S. (2001). Dimensions of Secondary School Effectiveness: Comparative Analyses Across Regions. School Effectiveness and School Improvement, 12, 3, 285-322.

Thrupp, M. (2001a). Sociological and Political Concerns about School Effectiveness Research: Time for a New Research Agenda. School Effectiveness and School Improvement, 12, 1, 7-40.

Thrupp, M. (2001b). Recent School Effectiveness Counter-critiques: Problems and Possibilities. British Educational Research Journal, 27, 4, 443-457.

Townsend, T. (2001a). The Background to This Set of Papers on the Impact of Two Decades of School Effectiveness Research. School Effectiveness and School Improvement, 12, 1, 3-5.

Townsend, T. (2001b). Satan or Saviour? An Analysis of Two Decades of School Effectiveness Research. School Effectiveness and School Improvement, 12, 1, 115-129.

Van den Noortgate, W., De Boeck, P., & Meulders, M. (Winter 2003). Cross-

Classification Multilevel Logistic Models in Psychometrics. Journal of Educational and Behavioral Statistics, 28, 4, 369-386.

Walford, G. (2002). Redefining School Effectiveness. Westminster Studies in Education, 25, 1, 47-58.

Waxman, H.C., Huang, S-Y, et al. (Sept/Oct. 1997). Classroom Process Differences in Inner-City Elementary Schools. Journal of Educational Research, 91, 1, 49-60.

Webster, B. J., & Fisher, D. L. (2000). Accounting for Variation in Science and Mathematics Achievement: A Multilevel Analysis of Australian Data Third International Mathematics and Science Study (TIMSS). School Effectiveness and School Improvement, 11, 3, 339-360.

Willmott, R. (1999). School Effectiveness Research: An Ideological Commitment? Journal of Philosophy of Education, 33, 2, 253-268.

Wong, K. C. (1996). How Effective Are Hong Kong Secondary Schools? Retrieved January 15, 2001, from http://www.hku.hk/rss/res_proj/31/31.html

Wyatt, T. (1996). School Effectiveness Research: Dead end, damp squib or smouldering fuse? Issues in Educational Research, 6, 1, 79-112.

Yang, M. & Goldstein, H. (December 1999). The Use of Assessment Data for School Improvement Purposes. Oxford Review of Education, 25, 4, 469-483.

Zehr, M. (October 2001). After Long Debate, Indiana Adopts Plan for Ranking Schools. Education Week, 21, 7.