Measuring School Effects Across Grades

Njora Hungi

FLINDERS UNIVERSITY INSTITUTE OF INTERNATIONAL EDUCATION
RESEARCH COLLECTION
NUMBER 6

STUDIES IN COMPARATIVE AND INTERNATIONAL EDUCATION
Number 6

Measuring School Effects Across Grades

NJORA HUNGI

FLINDERS UNIVERSITY
INSTITUTE OF INTERNATIONAL EDUCATION

Title: Measuring School Effects Across Grades
Series: Flinders University Institute of International Education Research Collection: Number 6
First Published: August 2003

Produced by the Flinders University Institute of International Education
Sturt Road, Bedford Park, Adelaide SA 5000
G.P.O. Box 2100, Adelaide 5001, AUSTRALIA
Email: FUIIE@flinders.edu.au
Website: http://wwwed.sturt.flinders.edu.au/fuiie

Edited by John P. Keeves and Jonathan Anderson
Designed by Katherine L. Dix
Published by Shannon Research Press, South Australia
ISBN: 1-920736-02-6

Preface

The general purpose of this study is to investigate the issue of the value-added components of the education provided across Grade 3 and Grade 5 in primary schools in South Australia and how these components could be measured. The data for this study were obtained from the Department for Education, Training and Employment (DETE) in South Australia. These data have been collected annually as student responses to Basic Skills Tests (BST) administered to Grades 3 and 5 students in government schools throughout South Australia since the inception of the Basic Skills Testing Program (BSTP) in 1995.

This study argues that, if primary schools in South Australia are to be assessed in terms of the value added to students' achievement over the two-year period, then it would be necessary to allow for the performance of the students, before the commencement of the period under review.

In this study, all the Grades 3 and 5 Basic Skill Tests from the six testing occasions (1995 to 2000) are calibrated separately using the Rasch model after which the concurrent equating method is used to link the tests to form two scales: one for Literacy and the other for Numeracy. The two scales are then used to obtain scores for every student from the six testing occasions at the Grades 3 and 5 levels. The hierarchical linear modelling technique is then employed to investigate the effects of student-level and school-level factors on achievement in Literacy and Numeracy. In addition, the hierarchical linear modelling technique is employed to compute the value-added score for each school involved in the study. The resulting value-added scores are examined for consistency across (i) subject areas, (ii) testing occasions, and (iii) categories of students.

The analyses undertaken in this study, at Grade 5 primary school level in South Australia yield the following findings:

- Achievement at Grade 3, age, gender, racial background, migrant status, transience and English spoken at home are among the important individual-level predictors of student performance in the Basic Skills Testing Program.
- Average socioeconomic status, location, mobility and absenteeism rates are among the important school-level predictors of student performance in the Basic Skills Testing Program.
- The variance between students within schools in terms of their achievement in numeracy and literacy is roughly around four times greater than the variance in performance between schools.
- In the models employed to estimate school effects, after taking into account student background characteristics and achievement in the Basic Skills Tests at

Grade 3, substantial variance (about 30 to 40 per cent) between the students is left unexplained. However, a very small (one to three per cent) amount of the variance available at the school-level is left unexplained.

- A considerable number of schools that show more than expected average levels of performance in numeracy are also likely to show more than expected average levels of performance in literacy. However, only a small number of schools that show more than an expected increase in performance over time in numeracy are likely to show more than an expected increase in performance over time in literacy.
- Only a small number of schools that are relatively effective for one cohort of students in numeracy (or in literacy) are likely to be relatively effective for another cohort of students.
- A vast majority of schools that record more than an expected average performance in numeracy (or literacy) for boys also record more than an expected average performance for girls in numeracy (or literacy).

This study shows that, within the South Australian situation, it is very difficult to identify effective or ineffective schools because the amount of variance left unexplained at the school-level is small. As a solution to this problem, this study demonstrates that it is more meaningful to identify effective or ineffective schools when the school effects are expressed in terms of years of learning that a student spends at school.

In addition, this study argues that, because a substantial amount of variance is left unexplained within the school, future research on school effectiveness of primary schools in South Australia should focus on what is happening within the classrooms of the schools, rather than between schools.

Acknowledgments

There are many to whom I am indebted. Foremost among them I want to thank Professor John P. Keeves and Professor J. Anderson, my supervisors, for their invaluable guidance and assistance in every way. They provided me not only with intellectual guidance but also with moral support and encouragement, without which this work would not have been possible. Thank you indeed for pushing me well beyond my limits!

I would like to express by sincere gratitude to those people who assisted me to get the data I used. Among them were: Mr. B. Schmidt, Mr. C. Payne, and Dr. S. Rothman. I would also like to thank Mr. D. Curtis and Dr. I. Darmawan who assisted me in proof reading some sections of this book.

My sincere gratitude goes to my friends Dr. A. Sivakumar, Dr. A. Tilahun and Mr. D. Curtis for their assistance, encouragement and stimulating discussions. Special thanks also go to Ms. M. Petrina for her friendship and emotional support.

I would also like to thank my parents and members of my family Florence Wangari, Ruth Mwihaki, Deogratius Thuku, Esther Wanjiku, Robert Hungi and Nicollet Wanjiku for their emotional support, patience, and sacrifice during my study.

Last but not least, I would like to thank the Flinders University of South Australia for the International Postgraduate Research Scholarship award that made this study possible.

Contents

Figures

Tables

1

Introduction

It does not need research findings to convince parents, teachers, and others concerned with schooling that there is a strong positive relationship between achievement in the basic skills of numeracy and literacy in the early years of schooling and future success in other school subjects as well as employment prospects. Findings from the Longitudinal Surveys of Australian Youth (LSAY) project indicate that students who achieve higher levels of numeracy and literacy during the compulsory years of schooling are more likely than students with lower levels of achievement to stay on to complete Grade 12 and proceed to institutions of higher learning (Marks et al., 2000). There is also research evidence that links achievement in numeracy and literacy to social outcomes such as community participation, engagement in lifelong learning, and health (Roberts and Fawcett, 1998).

Concerns that some children may fail to achieve an adequate level of competence in the basic skills of numeracy and literacy before they leave school at the end of the period of compulsory schooling are not new in Australia. Indeed, these concerns led to the first nation-wide basic skills survey in Australia in 1975, which was conducted by the Australian Council for Educational Research (ACER) following a request by the House of Representatives Select Committee on Specific Learning Difficulties (Keeves and Bourke, 1976). That survey called the Australian Studies in School Performance (ASSP) and it involved both primary and secondary levels, included schools from all States and Territories and from the government, Catholic, and independent school systems (Keeves, Matthews and Bourke, 1978).

Because of different school entry and grade promotion policies in different parts of Australia, little comparability existed between grade levels among States, and therefore sampling by age was chosen rather than sampling by grade in the ASSP survey (Keeves and Bourke, 1976). In addition, it was considered desirable to target students at an age level where the basic skills associated with reading, writing, and number work were likely to have been mastered and where a student possessing these skills was able to continue with learning in school with some degree of autonomy. Furthermore, it was considered important to target students at an age level just prior to the end of the period of compulsory education where all members of the age group

would still be at school. Consequently, the age levels selected for the ASSP survey were the 10- and 14-year-old levels.

The instruments employed to collect data in that ASSP survey were questionnaires and tests. The questionnaires focused mainly on capturing information concerning student background such as language spoken in the home, race, ethnicity, gender, need for remedial teaching, physical disabilities, learning problems, and location of residence. The tests aimed at measuring levels of competence in the basic skills of reading, writing and numeration. Some common (anchor) items were included in both the 10- and 14-year-old tests with the purpose of comparing performance between the two age groups.

The results from the 1975 ASSP survey indicated that a relatively small proportion of students failed to reach mastery levels in numeracy and literacy at both the 10- and 14-year-old levels. However, Keeves, Matthews and Bourke (1978) reported that this small proportion was considered to be meaningful in terms of absolute numbers of students across Australia. That survey also provided evidence that strongly suggested that some differences in performance between student groups on the sub-tests could probably have been due to factors such as the student's gender, race, ethnicity, location of residence, and language spoken in the home. There was also the general observation that performance of the 14-year-old students on the anchor items was noticeably higher than that of 10-year-old students on those items, suggesting growth in achievement between these two age groups.

Following the 1975 ASSP findings, the Parliament of the Commonwealth of Australia published the report of the House of Representatives Select Committee on Specific Learning Difficulties (1976) titled *Learning Difficulties in Children and Adults*. That report listed several recommendations on how to tackle the problem in Australian schools associated with the children failing to reach adequate levels in numeracy and literacy. Among the recommendations were: (a) the need for determination of learning difficulties by nation-wide surveys at regular intervals, (b) the addition of other age cohorts to the survey samples, and (c) development of the survey to include criteria that measured other competencies such as language and oral skills.

A second ASSP survey was carried out in 1980. From the second survey, it was evident that sizeable proportions of learners at both 10- and 14-year old levels were still failing to attain mastery skills in literacy and numeracy as had been found in the first survey (Bourke et al., 1981). This was despite the fact that there was a considerable improvement in performance compared to the performance in the first survey.

Unfortunately, the ASSP surveys were abandoned apparently due to opposition from teachers (Masters, 1994; p.16) and "only Tasmania continued the planned cycle of literacy and numeracy assessments extending the testing to all 10- and 14-year-olds". Nevertheless, the concerns of parents to obtain information on their children's levels of performance in the areas of learning associated with the basic skills of numeracy and literacy have during the past decade resulted in the introduction and maintenance of testing programs in most States and Territories in Australia.

In South Australia, the testing program known as the Basic Skills Testing Program (BSTP) was introduced in 1995 following a trial program in 41 schools in 1994. The program has continued every year since its inception, and it targets every Grade 3 and every Grade 5 student in all government schools throughout South Australia. In 2000, the BSTP was extended to include Grade 7 students.

The major purpose of the BSTP is to identify students having difficulties in areas of numeracy and literacy. It is hoped that early intervention strategies would help those

students identified as having difficulties to gain the necessary numeracy and literacy skills. Consequently, each student taking part in the BSTP is given an individual report. This report indicates items where the student's answers are correct, items where the answers are wrong and the learner band levels of performance on each sub-test. Each participating school is also given a report, which provides a summary of the performance of the students in that school on the tests.

Problem and its context

The inception of the BSTP in South Australia marked the beginning of ongoing heated debate between proponents and opponents of the program regarding the worth of the program. A majority of those opposed to the BSTP are mainly teachers, while the advocates of the program are parents and politicians within the State Government.

The critics of the BSTP mainly argue that the program is unnecessary because the information that is obtained from the program about levels of achievement of the individual student is in no way superior to what teachers can gather when teaching, based on their professional training and experiences. In addition, the opponents argue that the BSTP could pressurize teachers and schools to reform, which may not necessarily be for the better. Specifically, the opponents argue that the BSTP has the potential of causing teachers to alter the content of their classroom instruction to match the tests. In other words, teachers could find themselves in a situation where they are compelled to focus on test-specific materials or teach numeracy and literacy everyday and neglect other subjects.

On the other side of the debate, proponents claim that the program provides useful feedback to parents, teachers and educational administrators, and that this feedback is necessary if weaker students are to be identified and assisted to acquire the necessary basic skills of numeracy and literacy. Thus, the supporters of the BSTP claim that the program is useful because it is a diagnostic tool, which can assist in planning and teaching at individual, class and school levels. Furthermore, they claim that the program may assure the public about the standards of literacy and numeracy in public schools.

Most parents who support the BSTP argue that, although they have faith in teachers' judgement, they nevertheless are interested to learn how their children are faring in these crucial areas of school learning from an independent source. Judging from the large numbers of students who have participated in the BSTP since its inception in 1995, it would appear that a vast majority of parents support the program. For example, in 1998, over 95 per cent of the target group participated in the BSTP and this percentage is roughly the same for the other testing occasions.

So far there has been no attempt to rank or to publish the performance of the schools based on the their students' scores from the BSTP. However, proponents of the program have been quoted in a wide range of print and electronic media as having claimed that the results from the program show that the levels of achievement of successive cohorts of students have continued to increase since the inception of the program; and that, the results from the BSTP have shown that schools have improved in their performance since the inception of the program in 1995. As it would be expected, these claims have brought a new twist to the debate: that of the potential role of the BSTP as an instrument for assessing the performance of primary schools in South Australia and the development of school league tables as in Britain. Of course, this is the main latent reason why teachers have been opposed to the program, that is, the results from the program could be used to rank schools and somehow to hold them

and their teachers accountable for their students' levels of achievement in numeracy and literacy.

On a broader context, it is generally agreed that it is justifiable to compare the performance of schools so long as such comparison is based on sound understanding of the schools' data and circumstances. Schools do not operate in isolation, but as units within a national system entrusted with the task of providing education to its citizens. In order to meet the national education goals, there are sets of national benchmarks to be attained in all schools. Consequently, a school should not consider its performance in isolation but in comparison with others. Silins and Murray-Harvey (1998) argue that:

> Schools are accountable to students, parents and, more widely, to the community. Communities have a right to know how their schools are performing. Academic performance is one aspect of a school's performance that can be measured. (Silins and Murray-Harvey, 1998; p.10)

Apart from academic performance, there are other aspects that can be used to identify an effective school. However, for primary schools, academic performance especially in numeracy and literacy is very important because achievement in these two subjects has been shown to be a key factor influencing later educational, employment and social outcomes (Roberts and Fawcett, 1998; Marks et al., 2000). Furthermore, it could be argued that academic performance by a school is highly likely to be associated with other desirable contributions of school to the community.

Purpose of the study

The purpose of this study is neither to dispute nor support one side or the other in the debate regarding the worth of the BSTP in South Australia. Neither is the purpose of this study to develop new methods of measuring school performance or dispute existing methods, but rather to investigate using existing methods how the performance of the public schools in South Australia could be measured based on students' scores from the BSTP. Thus, the purpose of this study is to bring some research information to the debate (especially with respect to performance of the schools) by developing a general model upon which school effects could be estimated. In other words, the general purpose of this study is to investigate the issue of the 'value added' components of the education provided in the South Australian public primary schools and how these components for numeracy and literacy could be measured based on the students' scores from the BSTP.

Aims of the study

The major aims of this study within the general investigation of value added by schools, are:

(a) to develop common scales for measuring achievement in the Basic Skills Tests across the Grades 3 and 5 primary school levels and across six testing occasions (1995 to 2000) in South Australia;

(b) to examine the achievement levels of the Grades 3 and 5 students in the Basic Skills Tests in South Australia;

(c) to examine changes in the numeracy and literacy achievement levels of Grade 5 students in South Australia;

(d) to develop multilevel models of student-level and school-level factors influencing numeracy and literacy achievement of Grade 5 students in South Australia; and

(e) to investigate the issues associated with measuring the value added components of the education provided in the South Australian primary schools, and how these measured components could be based on Grade 5 students' scores from the Basic Skills Tests.

Meaning of the term 'value added'

The terms 'school effect' and 'value added' are used interchangeably in most school effectiveness studies. Hill (1996; p. 7) has defined 'value added' as a measure that indicates "the educational value that the school adds over and above that which could be given by the backgrounds and the prior attainment of the students within school". Another suitable definition is provided by McPherson (1993; p. 1); "a school's 'added value' is the boost it gives to a child's previous level of attainment". Here, the term attainment is used by McPherson to mean achievement.

A major British study known as Value Added National Project defined 'value added' as the progress schools help pupils to make relative to the their starting point (Fitz-Gibbon, 1997; Saunders, 1999). Thus, school effect or a school's value added component can appropriately be seen as the unexpected gain a school provides to its students.

Importance of value added measures

The identification of an adequate measure for comparison of schools in their effectiveness is a problem that has for a long time puzzled researchers (Bryk and Raudenbush, 1987; Hattie, 1992). The conservative comparisons of the effectiveness of schools using unadjusted average test scores of students does not appeal to researchers because such comparisons are "highly flawed even though derived from valid assessments" (Meyer, 1996; p.198). Comparison of schools using unadjusted average scores is argued to be inappropriate because students are not allocated to schools at random, neither are schools located in areas with similar neighbourhood characteristics, nor are all schools of the same size. As a result, schools differ in their student intakes as assessed by such characteristics as prior achievement, socioeconomic background, racial background, and ethnicity. Past studies in Australia and internationally have ascertained that student-level factors such as level of achievement at entry, family background, and school-level factors such as locality and school type may affect students' level of achievement (e.g. Husén, 1967; Comber and Keeves, 1973; Keeves, 1975; Postlethwaite and Wiley, 1991).

Consequently, researchers argue that it would be misleading to compare schools using average scores without adjusting the scores for the differences between schools for at least the achievement level of their students at entry (e.g. McPherson, 1993; Yang et al., 1999). The adjustment of the scores for the differences between schools enables the boost (value added) that each school provides to its students' achievement to be more apparent. Thus, value added measures are intended to allow fairer comparison between schools.

In addition, it is argued that value added measures provide useful information that can be used by a school in improving the performance of students, staff and the school as a whole. Furthermore, the information is useful in assisting those in appropriate positions (or interested in) to pass judgement on schools based upon a good

understanding of the schools' circumstances. Those in a position to judge schools include parents choosing schools for their children and funding agents evaluating school practice (Raudenbush and Willms, 1995).

It is important to bear in mind that comparisons between schools based on value added measures are relative ones; "that is, they position each institution in relation to other institutions with which they are being compared" (Goldstein, 1997; p. 372). Thus, the use of the descriptive terms ‘effective’ or ‘ineffective’ could be misleading because, based on some absolute criterion, all the schools being compared could be performing poorly or all the schools could be performing well (Coe and Fitz-Gibbon, 1998).

Specific research questions

This study aims at addressing most of the important issues in the measurement of value added components. It is considered essential to address these issues in order to provide a technically sound solution[1] to the problem of measurement of the value-added components of the education provided in the South Australian primary schools. Consequently, the study address 24 specific research questions, which are presented below. These research questions are based on the data available for this study (see Chapter 3).

1. Is there adequate fit of the Rasch model to the Grades 3 and 5 items?
2. How do the average item difficulties of the Grades 3 and 5 tests compare across testing occasions?
3. Can the numeracy items for 1995 to 2000 tests for Grades 3 and 5 be brought to a common scale?
4. Can the literacy items for 1995 to 2000 tests for Grades 3 and 5 be brought to a common scale?
5. Has the level of performance in numeracy (or literacy) at Grade 5 changed significantly over time?
6. What is the average growth in numeracy and literacy achievement between Grades 3 and 5 levels?
7. What student-level factors influence numeracy (or literacy) achievement?
8. What school-level factors influence numeracy (or literacy) achievement?
9. What cross-level interaction effects influence numeracy (or literacy) achievement?
10. What amounts of variance are available at the student-level, school-level and occasion-level?
11. What percentages of variance in student scores in numeracy and literacy at Grade 5 are explained by Prior Achievement (that is, achievement at Grade 3) alone?
12. What percentages of variance in student scores in numeracy and literacy at Grade 5 do the predictor variables in the final two-level and three-level models explain?
13. What percentages of variance in student scores in numeracy and literacy at Grade 5 are explained in the models employed to estimate school effects?

[1] For example, based on McPherson's "explicit theory of good standing" (1996; p.1) and Meyer's "attributes of acceptable and valid school effectiveness indicators" (1996; p.178).

14. What percentages of variance in student scores in numeracy and literacy at Grade 5 are left unexplained at the student-level in the models employed to estimate school effects?
15. What percentages of variance in student scores in numeracy and literacy at Grade 5 are left unexplained at the school-level in the models employed to estimate school effects?
16. How reliably are the school effects estimated?
17. Can a stability index be calculated to compare the stability of the various types of school effects over time?
18. Based on value added scores, is the rank order of the schools, using all the students who could be matched, greatly different from the rank order of the schools using only those students who could be matched in the same school?
19. Are schools that are identified as relatively effective based on one type of school effect also identified as relatively effective based on a different type of school effect?
20. Do schools that show more than expected average levels of performance also show more than expected increases in performance over time?
21. Are schools that are relatively effective in numeracy also relatively effective in literacy?
22. Are schools that are relatively effective for one cohort of students also relatively effective for other cohorts of students?
23. Are schools that are relatively effective in numeracy for boys also relatively effective for girls?
24. Are schools that are relatively effective in literacy for girls also relatively effective for boys?

The answers to the above research questions are provided in Chapter 12.

Significance of the study

This study aims to investigate the issue of the value added components of the education provided in the South Australian primary schools and how these components could be measured considering factors that influence student performance. Therefore, this study should bring important research information to the ongoing debate regarding the potential usefulness of the BSTP in assessing the performance of public schools in South Australia. In addition, this study should make a significant contribution to the following areas of knowledge concerning effectiveness of each primary school in this State:

(a) overall school effectiveness;

(b) stability in school effectiveness across successive cohorts of students;

(c) change in school effectiveness over time;

(d) consistency in school effectiveness across subjects; and

(e) consistency in school effectiveness across various categories of students considering student's characteristics, such as: gender, English speaking background and racial background.

The study also involves development of common scales for numeracy and literacy across the different testing occasions of the BSTP since its inception in South Australia in 1995. Thus, the study should contribute significantly to knowledge about change in performance of students in basic skills of numeracy and literacy in South Australia over time. It should be possible to establish whether the performance of students is improving, being maintained or deteriorating.

Furthermore, this study is the first of its kind in the study of basic skills in South Australia. It will be an important milestone in educational research in this State and in Australia especially since there is no comprehensive literature and research studies on issues relating to measurement of the value added component of schools both in South Australia and in Australia.

Moreover, on a broad level, the knowledge gained from this study should be useful to those concerned with school performance issues in other States and countries around the world.

Limitations of the study

The current study uses secondary data collected by the Department of Education Training and Employment (DETE) in South Australia and Department of School Education in New South Wales. The common problem that faces any study based on secondary data analysis, is the fact that the research is restricted to the variables present in the data. Therefore, the design of the current study is limited to the existing variables. Nevertheless, a careful examination of the available data was carried out before commencing the study. Generally, the examination of the available data revealed that the data contained most of the important information needed to form variables that had research backing as possible predictors of student achievement. However, at the student-level, the data lacked information on socioeconomic status and absenteeism but some of this information was available at the school-level.

School effectiveness studies are often criticized for using variables as proxies for unmeasured characteristics with which they are associated. For example, Coe and Fitz-Gibbon (1998) argue against the use of variable such as 'sex' or 'ethnic origin' unless there is supportive evidence to show that the effects result from purely biological differences, or from unfair discrimination. For the variable 'sex', Coe and Fitz-Gibbon argue that if gender differences in mathematics achievement are attributed to a spatial visualization factor (which is stronger in males than in females), it would be more appropriate to measure this factor and include it in the analyses rather than stereotype all girls. However, for this study, no additional data can be obtained, and therefore variables such as 'sex' and 'racial background' are used but it is recognized that such variables are crude proxies.

Another limitation of this study is lack of data at the class-level, which means performance of classes or teachers within a school cannot be examined. The issue of the school as the unit of analysis is addressed in more detail in the next sub-section.

School versus class as the unit of analysis

It is generally argued that within schools there are some classes (or teachers) that are more effective than others are. Indeed, results from school effectiveness studies that have taken into account differences among classes (or teachers) within schools have indicated that those differences outweigh the differences between schools (e.g. Creemers and Reezigt, 1996; Einsiedler and Treinies, 1997; Fitz-Gibbon, 1991 & 1997; Kyriakides et al., 2000).

At the primary school level, Bressoux (1995) argues that the rate of learning to read by a child depends mostly on the class the child is in as well as the teaching methods used and less on the school itself. Bressoux associates the differences between schools to one or two classes which tends to raise (or lower) the level within that school and not a global progression of all classes together.

Work by Kyriakides et al. (2000) reported that 13.8 per cent of the variance in mathematics achievement among final year primary pupils in Cyprus lay at the class-level, which was substantially higher than the amount of variance reported at the school-level (8.5 per cent).

At the secondary school level, in the United Kingdom, Fitz-Gibbon (1991; p.81) reported that analyses conducted in the first year of the A-level Information System (ALIS) in 1983 indicated that "classes-within-schools were as different as different schools". In addition, Fitz-Gibbon (1997) reported that analyses of data from the Value Added National Project (VANP) showed that up to 42 per cent of the student-level variance on the external examination at age 16 years was associated with teachers. However, the students in the VANP study were not randomly allocated to the different teachers (that is, students may have been placed in classes according to their ability levels) and it was likely the actual proportion of variance attributed to teacher effect could be lower than observed.

Similarly, a number of research studies on instructional techniques of secondary school teachers in the Netherlands have reported that the differences between teachers within schools were much larger than differences between schools (Bosker and Akkermans, 1994; Heyl, 1996). However, it should be borne in mind that, just as in the United Kingdom, secondary schools in the Netherlands were characterized by streaming of students and these might have provided an inappropriate and inadequate picture of the variance between classes in these studies.

In Australia, a study by Webster and Fisher (2000) using data collected as part of TIMSS reported that 33.8 per cent and 7.6 per cent of the variance in mathematics achievement lay at the class-level and school-level respectively. However, work by Luyten and de Jong (1998) in the Netherlands reported little difference in student achievement between parallel classes taught by different teachers, and suggested that the differences found across classes might be due only to loose internal conditions. Moreover, a study by Hill and Goldstein (1998) found that any observed large within-school-between-class differences tend to cancel each other out over a period of two years, and that eventually the differences between schools emerge as relatively large.

Notwithstanding these findings, as evidence became available that differences between classes could outweigh the differences between schools, a number of commentators started to argue that research studies into student progress should focus on class effects rather than school effects. However, researchers in this field have argued that for practical purposes it is more useful to regard the school rather than the class as the unit of analysis in school effectiveness studies (e.g. Witte and Walsh, 1990; Hill, 1996; Silins and Murray-Harvey, 1998; Teddlie et al., 2001). In justifying the school as the unit of analysis, Hill argues that:

> Schools constitute a natural and a relatively discrete unit of analysis and it is primarily at this level that issues of parental choice arise. Many if not most key decisions concerning resources and programs are made at the level of the school. (Hill, 1996; p.9)

It is logical to expect schools to monitor and control most of what goes on in their classes. Thus, if the aim is to hold schools accountable for their performance, it is reasonable to consider the school as the unit of analysis in school effectiveness

studies. However, if the aim is to monitor classes (or teachers) within schools, then an analysis with classes as the unit of interest is warranted (see Teddlie et al., 2001; pp. 96-100).

For the South Australian situation, it is worth noting that the teachers' union in this State is opposed to the use of the class as the unit of analysis. The union argues that such analysis could put too much pressure on teachers, and results from such an analysis have the potential of being misused to victimize teachers.

Structure of the book

The first two chapters provide a general setting for the study. The background to the study is given at the beginning of this chapter by tracing the importance of achievement in numeracy and literacy and presenting a short history of the BSTP in South Australia. This chapter also defines the problem, the general purpose and aims of the study, the significance and limitations of the study, introduces the research questions and explains why school rather than class is used as the unit of analysis in this study.

In Chapter 2, literature reviews of issues related to test theories, equating of tests and school effectiveness research that are of interest to the current study are presented. The chapter provides background information about the main concepts of the theories employed to analyze the tests in this study; namely, classical test theory and item response theory. It also provides background information about the equating methods employed to bring the tests onto common scales and gives summaries of what past studies have said or found regarding school effectiveness issues that are of interest in this study.

Chapters 3, 4 and 5 introduce the data, variables, design, and models employed in this study. Chapter 3 describes the instruments used to collect data, the data sets available for this study and the construction of variables from these data sets. In Chapter 4, the methods of data analysis and the computer packages employed in this study are described. Chapter 5 describes the general design and models employed in this study to answer the research questions raised above.

Chapter 6 describes the steps followed to equate all the Grade 3 and Grade 5 tests from the six occasions (1995 to 2000) to construct common scales: one for numeracy and the other for literacy. The common scales described in Chapter 6 are used in the construction of achievement related variables, (that is, student scores at Grades 3 and 5 in the Basic Skills Tests), which are used in subsequent analyses in this study.

Chapters 7 to 11 detail multilevel techniques employed to tease out the factors influencing student achievement and to estimate different types of school effects. Chapters 7 and 8 report on analyses carried out to examine factors influencing achievement in numeracy and literacy among Grades 3 and 5 primary school students in South Australia. Chapter 7 focuses on two-level analyses while Chapter 8 focuses on three-level analyses. Chapters 9 and 10 describe an approach to examining performance of primary schools in South Australia over time using the scores from the BSTP on several cohorts of students and based on a longitudinal multilevel structure. For both numeracy and literacy, the longitudinal structure mentioned above is employed to estimate different types of school effects or value-added scores. The resulting value-added scores are examined for consistency across subject areas, and across cohorts of students.

Chapter 11 reports on analyses carried out to examine the consistency of school effects across gender groups. The longitudinal multilevel structure mentioned above is

employed to estimate indices of individual school effectiveness for boys and girls for numeracy and literacy using two approaches. Chapter 12 gives the main conclusions from the findings of the study by providing answers to the research questions. Implications of the findings of the study for theory, practice and further research are also given in this chapter.

2

Literature Review

In this chapter, reviews of issues related to test theories, equating of tests and school effectiveness research that are of interest to the current study are presented. These reviews are presented here in order to provide background information of the methods and theories employed in this study and, therefore, facilitate the understanding of the discussions and analyses that are presented in subsequent chapters.

The structure of this chapter is as follows. The first two sections provide background information about the main concepts of the theories employed to analyze the tests in this study; namely, classical test theory and item response theory. The third section provides background information about the equating methods employed to bring the tests onto common scales. The fourth to the eighth sections focus on what past studies have said or found regarding school effectiveness issues that are of interest in this study: namely, (a) issues for research in school effectiveness, (b) school effect indices, and (c) multilevel modelling in school effectiveness research (SER).

Classical test theory

Classical test theory (CTT) proposes that there is "a linear relationship between a person's observed number-correct test score and the error-free true score that it estimates" (Weiss and Yoes, 1991; p.70).

Within this theory, a true score plus an error gives a person's score on a test and the model employed is expressed mathematically as:

Observed Score = True Score + Error Equation 2.1

(Weiss and Yoes, 1991; p.70)

The true score and the error are assumed to be un-correlated (Lord, 1980; Keats, 1997).

In CTT, an individual's scores are based on the number of items to which they respond correctly (Weiss and Yoes, 1991). In other words, a person's total score on a

test is equal to the number of items the person answered correctly or a function thereof. Scores calculated in this manner are referred to as number-correct scores.

Essentially, classical test theory requires that items in the test being analyzed measure a common variable or are unidimensional (Lord, 1980). Consequently, it is necessary to check that the test items satisfy this unidimensionality requirement before estimation can proceed. Keeves and Alagumalai (1999: p.10) contend that in CTT, "item analysis procedures are employed, and a reliability index is calculated in order to support the meaningfulness of a total score". However, if there are doubts regarding the underlying dimensionality of a test, the unidimensionality requirement of the test should be established using confirmatory factor analysis of an item intercorrelation matrix (Lord, 1980; Marsh and Hocevar, 1983; Hattie, 1985; Weiss and Yoes, 1991; Vijver and Poortinga, 1991; Spearritt, 1997).

It should be emphasized that unidimensionality does not imply a single factor because as Hambleton pointed out:

> What is required for the assumption of unidimensionality to be met to a satisfactory extent by a set of test data is a *dominant* component or factor. (Hambleton, 1989; p.150)

Furthermore, Bejar (1983) clarified that as long as a set of items function in unison, that is, the same psychological processes affect the performance on each item in the same form, unidimensionality will hold. Hence, other factors could be present but as long as a set of test data contains a dominant component, the test can be regarded as having met the requirements of unidimensionality.

Classical test theory has three major shortcomings. First, the values of the estimated parameters of the test items (item difficulty and item discrimination) depend on the particular sample of students to whom the items were administered (Osterlind, 1983; Hambleton and Swaminathan, 1985; Wright, 1988; Hambleton, 1989; Weiss and Yoes, 1991, Barnard, 1999).

Second, critics argue that number-correct scores are dependent on the difficulties of the items selected for use in the test (Weiss and Yoes, 1991). That is, the case estimates or scores are dependent upon the sample of items in the test. For example, a student could obtain a high score if given Test-A containing simple items, while the same student could obtain a low score if given Test-B measuring the same attribute as Test-A but containing difficult items.

Finally, the concept of reliability as defined in CTT is also dependent upon the particular sample of students involved in the total score distribution (Weiss and Yoes, 1991). This is because, under CTT, calculation of reliability involves total score variance, which depends on the sample of students involved in the test. In addition, Hambleton (1989) argues that the main problem concerning the concept of test reliability stems from the fact that the concept is generally defined in terms of parallel-forms of a test which are difficult to achieve in practice.

Critics argue that because of the above three shortcomings of CTT the theory has failed to provide satisfactory solutions to many testing problems (Hambleton and Swaminathan, 1985; and Weiss and Yoes, 1991). Weiss and Yoes, add that “it is at least partly in response to these recognized inadequacies of CTT that IRT[2] was developed” (p.70). However, Lord (1980) has argued that IRT supplements rather than contradicts CTT. Furthermore, Barnard (1999) has argued that the results obtained from a CTT based item analysis can yield useful information in finding flaws

[2] item response theory

in items and sometimes guiding the test developer towards choosing an appropriate IRT model.

Item response theory

In some scholarly writings, item response theory (IRT) is referred to as the 'latent trait model' (Wilcox, 1988), 'latent trait measurement model' (Douglas, G., 1988), or 'item characteristic curve theory' (Osterlind, 1983). This test theory proposes that the relationship between a student's performance and the probability that the student will answer an item correctly can be described using a mathematical function (Lawley 1942; Stocking, 1999). The mathematical function is referred to as the item response function (IRF) or item characteristic curve (ICC) and it provides the probability of examinees answering an item correctly for examinees at different points on the proficiency scale (Lord, 1980; Hambleton and Swaminathan, 1985; Dorans, 1990; Hambleton et al., 1991; Embretson, 1999).

Proponents of IRT argue that, unlike classical test theory, item parameters obtained using IRT are independent of the group of students used from the population of students for whom the test was designed (Wright, 1988; Hambleton, 1989; Weiss and Yoes, 1991; Kline, 1993). In addition, the supporters of IRT argue that the theory permits students' performance to be estimated independently of the test items used, and therefore, provides some basis for determining how an examinee might perform when confronted with a test item (Hambleton and Swaminathan, 1985; Weiss and Yoes, 1991; Stocking, 1997). Moreover, the proponents of the model claim that IRT provides scores on an interval scale that extends indefinitely above and below a zero score corresponding to the average difficulty level of the items.

Weiss and Yoes, (1991) have noted that ICC can assume various shapes depending on the parameters included in the mathematical equation used to describe the item, that is, the item response function. The number of parameters included in the mathematical form of the IRF has resulted in three major measurement models within the item response theory. These three IRT models are simply defined as the one-, two-, and three-parameter models and are respectively expressed mathematically by Weiss and Yoes as follows:

$$P(u = 1/\theta) = [1 + e^{-D(\theta - b)}]^{-1} \qquad \text{Equation 2.2}$$

$$P(u = 1/\theta) = [1 + e^{-Da(\theta - b)}]^{-1} \qquad \text{Equation 2.3}$$

$$P(u = 1/\theta) = c + (1 - c)\,[1 + e^{-Da(\theta - b)}]^{-1} \qquad \text{Equation 2.4}$$

where

$P(u = 1/\theta)$ represents the probability of a correct response to a given item by an examinee with ability θ, and b is the difficulty of the item,

D is a constant equal to 1.7,

a is the discrimination parameter for the item; and

c is a chance scoring (or guessing) parameter.

(Weiss and Yoes, 1991; pp.77-8)

The one-parameter logistic model (see Equation 2.2) is popularly known as the Rasch model after its developer, Georg Rasch, in 1960. In this model, the probability of a student answering an item correctly is defined as a function of the student's ability and

the difficulty of the item, without taking into consideration either the item discrimination parameter or a guessing factor associated with each item (Lord, 1982).

Specifically, the Rasch model requires that; (a) all items in a test have equal discriminating power, and (b) guessing in a test is minimal (Scheuneman, 1979; Lord, 1980; Hambleton and Swaminathan, 1985). Some psychometricians have been especially doubtful about the appropriateness of these two requirements. Some argue that guessing plays a considerable part in answering multiple-choice items (Choppin, 1992 & 1997; Rogers, 1997) and that achievement test items differ in the degree to which they correlate with the underlying trait and therefore it is not appropriate to assume uniformity in discrimination power of the items in a test (Hambleton and Swaminathan, 1985; Kline, 1993; Stocking, 1999). However, the proponents of the Rasch model argue that guessing is a characteristic of only a few individuals and not the items and that the model is fairly robust with respect to departures of model requirements normally observed in actual test data (Hambleton and Swaminathan, 1985; Skaggs and Lissitz, 1986).

McNamara (1996) has pointed out that the primary advantage of the Rasch model is that it makes possible estimates of item difficulty which are independent of the ability of persons, and estimates of person ability which are independent of the difficulty of items. This is because the model employs the parameters of a person's ability and item difficulty to estimate the probabilities of the correct response of the person to the item. Therefore, the items may be used to obtain accurate scores for students regardless of their level of performance because the model can distinguish between level of performance and item difficulty. In addition, the Rasch model has fewer item parameters, which makes it easier to work with compared to other IRT models (Hambleton, 1989).

There are several forms of the Rasch model. Wright (1988), and Andrich and Masters (1988) have provided general introductions to the various forms of the Rasch model. A volume edited by Fischer and Molenaar (1995) provides detailed accounts of the various forms of the Rasch model. Several entries in Masters and Keeves (1999), and a recent volume by Bond and Fox (2001), also provide more detailed accounts of the model. A highly technical account of the Rasch model can be found in Allerup (1997).

The two-parameter logistic model (see Equation 2.3) proposes that a student's probability of answering an item correctly is a function of the student's ability and item difficulty after taking into consideration the item discrimination, but not the chance or the guessing factor associated with each item. Thus, the two-parameter model allows items in a test to have varying discriminating power, but requires that guessing in a test is minimal. Therefore, the two-parameter model overcomes the equal-discriminating-power problem associated with the Rasch model but not the guessing problem. However, the two-parameter model is more complex than the Rasch model. Furthermore, because the discriminating power of the item is allowed to vary, the two-parameter model does not enable item parameters to be estimated sample free, and thus does not satisfy the requirements associated with measurement.

The three-parameter logistic model (see Equation 2.4) proposes that a student's probability of answering an item correctly is a function of the student's ability and item difficulty after taking into consideration the item discrimination as well as the guessing factor associated with each item. This model allows items in a test to have varying discriminating power and accommodates guessing as a characteristic of the item. Therefore, the model overcomes the problems of equal discriminating power as well as certain aspects of the guessing problems associated with the Rasch model. However, in order to fit the model to a set of data, a very large number of cases is required so as to obtain convergence and stable estimations (Kolen, 1994). This model

is also more complex compared to the one- and the two-parameter models. Furthermore, because the three-parameter model, like the two-parameter model, allows the discriminating power of the items to vary, the three-parameter model does not permit item parameters to be estimated sample free, and consequently does not satisfy the conditions for measurement.

It should be noted that all the three IRT models like CTT require that there is a single underlying latent trait that is being measured, that is, a single dimensionality of the latent space (Weiss and Yoes, 1991). Keeves and Alagumalai (1999) have argued that the IRT requires more stringent tests of unidimensionality compared to CTT. Consequently, they have proposed that a check of the fit of the items be employed to establish whether or not items in a test meet the dimensionality condition of the IRT model chosen. Moreover, if the Rasch model is employed, it is necessary that the person used in the calibration of a scale should also satisfy the requirements of the model.

In this study, the one-parameter IRT model, or the Rasch measurement model, is used because it is the most robust of the different methods available, and in addition, it is the only model that has strong measurement proportions (Sontag, 1984).

Test equating

Test equating describes a process that enables test developers and users to compare scores from different forms of a test (Stocking, 1997; Woldbeck, 1998). The process involves developing a conversion system that can then be used to change the units of one form of a test to the units of another form of the test. The conversion of the scores derived from the different forms of the test allows the scores to be used interchangeably (Petersen et al., 1989). Hence, after successful equating, examinees are expected to earn the same score regardless of the test form administered (Angoff, 1982; Lord and Stocking, 1988; Kolen, 1994).

Lord (1980) has argued that mathematically it is possible to 'equate' any two tests by mere manipulation of the scores of the tests. However, he notes that in reality test equating can only be seen as meaningful when certain equating conditions are met. The reason for this is apparent especially if the prime purpose of equating tests is to establish as nearly as possible, an effective equivalence between scores on the two tests. It should also be remembered that after successful equating, the two equated tests are therefore considered to be a measure of the same attribute with an equivalent degree of precision. It is in this connection that Lord (1980) proposed that scores on test X and test Y could only be equated if four conditions were met. Petersen et al. have summarized these four conditions for the equating of tests as follows.

1. Same ability - the tests must both be a measure of the same characteristic (latent trait, ability, or skill).
2. Equity - for every group of examinees of identical ability, the conditional frequency distribution of scores on test *Y*, after transformation, is the same as the conditional frequency distribution of scores on test *X*.
3. Population invariance - the transformation is the same regardless of the group from which it is derived.
4. Symmetry - the transformation is invertible, that is, the mapping of scores from form *X* to form *Y* is the same as the mapping of scores from form *Y* to form *X*. (Petersen et al., 1989; p.242)

Petersen et al. note that in reality it is unlikely that all the four mentioned above conditions would be met. However, they add, "there seems to be general agreement among practitioners that the equated scores should satisfy the population-invariance and symmetry conditions" (p.242). Since in practice it is unlikely to meet all the four conditions of equating, Petersen et al. conclude that there is probably no equating method that would produce truly equivalent scores on two forms of the same test. Nevertheless, they emphasize that since test scores can have important consequences for students, "an approximate equating of scores on two forms of a test will generally be more equitable to the examinees than no equating at all" (Petersen et al., p.243).

Most of the emphasis in the test equating literature focus on the dimensionality of the tests being equated (Engelhard, 1980; Hutten, 1980; Holmes, 1982; Bogan and Yen, 1983; Skaggs and Lissitz, 1986; Camilli, 1993; Kolen, 1997; Bolt, 1999). In particular, a majority of the equating studies have expressed the need for the tests being equated to be measuring the same characteristic and to be unidimensional. Bogan and Yen (1983) demonstrated that multidimensional tests usually yield worse equating than unidimensional tests, particularly when the tests being equated differ in difficulty. However, another study carried out by Dorans and Kingston (1985) indicated that although violations of unidimensionality in IRT equating may have an impact on equating, the effect may not be substantial.

It should be noted that a study of the factor structure and the scaling (calibration) characteristics of the BST, carried out by Hungi (1997) using the 1995 BSTP data, established that it is appropriate to equate and to calculate scores for:

(a) General Performance, Literacy, and Numeracy;

(b) Literacy, Language, and Reading; and

(c) Numeracy but only for curriculum purposes on Number, Measurement and Space, since in general the scores for those three factors should not differ greatly.

Thus, in this study it is appropriate to carry out the equating of the tests so as to calculate students' scores based on a single Literacy scale and a single Numeracy scale. Furthermore, since the Rasch model is preferred in this study, a check of the fit of the model to the data after successful equating should establish whether or not the items in a test meet the dimensionality condition of the model (Lake, 1998; Mohandas, 1999; Banerji, 2000; Waugh, 2001).

Forms of test equating

There are two general forms of test equating, namely 'vertical equating' and 'horizontal equating' (Hambleton and Swaminathan, 1985; Weiss and Yoes, 1991; Keeves, 1992a). These two equating procedures are discussed separately in the following sub-sections.

Vertical equating

Vertical equating is applicable where the tests to be equated are at different levels of difficulty and the ability distributions of the examinees are different (Woldbeck, 1998). This type of equating allows the scores of students at different levels to be compared and allows for the assessment of an individual student's development over time (Petersen et al., 1989; Kolen, 1994; Kolen, 1997). Hambleton and Swaminathan (1985), and Weiss and Yoes (1991) note, that in a vertical equating situation, the objective is to construct a single scale that would permit comparison of the abilities of the examinees at different levels (for example, at different grades). The tests

administered at the various levels (grades) are not multiple forms of one particular test and are obviously at different levels of difficulty. However, in order to equate the two tests, some common (also called 'anchor') items are included in both tests.

In this study, vertical equating is employed to link the Grade 3 to the Grade 5 tests on the same testing occasions. This vertical equating across the two grade levels is achieved through the use of common items included in the tests administered to the Grades 3 and 5 students on the same testing occasion.

Horizontal equating

Horizontal equating is a form of test equating that is applicable where the tests to be equated are at a comparable level of difficulty and the ability distribution of the examinees taking the test are similar (Hambleton and Swaminathan, 1985; Weiss and Yoes, 1991; Keeves, 1992a; Woldbeck, 1998). This form of equating is appropriate when multiple forms of a test are required for security and other reasons (Jaeger, 1980 & 1981; Cook and Eignor, 1991; Kolen, 1994). The various forms of the test are not identical but are expected to be parallel. It is also expected that the distribution of the abilities of the examinees to whom these forms are administered are approximately equal (Hambleton and Swaminathan, 1985). Ideally horizontal equating is aimed at making the two tests become true measures of the same psychological function with the same degree of accuracy and precision regardless of the group of examinees under consideration.

There are two main approaches to horizontal equating. One approach involves the use of common items that are included in the two forms of the tests to be equated and the other approach involves the use of common persons who take portions of the different forms of the test being equated. If the question papers for different forms of the test are allowed to circulate freely in the society, it is difficult to guarantee the security of the test regardless of the horizontal equating approach preferred and this could generally be seen as a major setback to horizontal equating. However, if all the question papers for the different forms of the test are collected and destroyed after each testing occasion, then the security of the test can be increased.

In this study, common persons horizontal equating approach is employed to equate the different forms of the tests that have been administered to the Grades 3 and 5 students since the inception of the BSTP in South Australia in 1995. This is because in the BSTP there are no items included on more than one formal testing occasion. The data necessary to link the six different forms of the test administered to South Australian students since the inception of the BSTP were obtained from New South Wales (NSW). These NSW equating data consist of groups of Grades 3 and 5 students who have taken the 1996 test (or another test directly linked to the 1996 test) as a trial test a week prior to taking the real test for that occasion. These data are described in more detail in Chapter 3.

Methods of test equating using the Rasch model

Past studies indicate that there are three different equating procedures that are commonly employed to equate tests with the Rasch model: (a) anchor item equating, (b) common item differences equating, and (c) concurrent equating.

In all the three procedures, the tests to be equated are first calibrated before equating is undertaken. The calibration involves the deletion of the misfitting items and sometimes the deletion of the misfitting persons. However, it should be noted that a study carried out by Phillips (1986) to investigate the effects of deletion of misfitting

persons in vertical equating using the Rasch model found no basis in the results for choosing between the two approaches (deleting or not deleting misfitting persons) in Rasch equating.

The next three sub-sections provide brief discussions of the three equating procedures mentioned above that are commonly employed to equate tests using the Rasch model.

Anchor item equating

An anchor item equating procedure involves anchoring the threshold values of the common items obtained from one of the tests in the calibration of the second test. For example, the threshold values of the common items in test-A would be anchored in test-B. The anchored items are then used to estimate the threshold values of the rest of the items in that test.

Experience shows that some of the anchored items acquire infit mean square (INFIT MNSQ) values outside the desired range especially when dealing with vertical equating. This is especially true with the more difficult common items. This raises the question of what should be done with such items. Deletion of the items may leave few common items for the estimation of the difficulty levels of the other items in the analysis. On the other hand, leaving such items in the analysis would attract criticism since, ideally, such items might not be measuring the same underlying attribute as the rest of the items in the test.

It should be noted that anchor item equating could be carried out with common persons instead of common items. In this case, the estimates of the common persons in one test are anchored and used to calculate the threshold values of the items in the second test.

Common item difference equating

A common item difference equating approach involves the computation of the mean of the differences of the threshold values of the common items in the two tests being equated. First, the threshold values of the common items in the first test are subtracted from the threshold values of the common items in the second test. Second, the differences are added and divided by the number of common items to obtain the average threshold difference between the two tests. In addition to estimating the mean difference, the error associated with that mean difference can also be estimated since the items are fixed as common items.

With the common item difference procedure, there is no direct interaction in the data of the two tests being equated. Hence, it is difficult to ascertain whether the items in the two tests being equated measure the same underlying attribute. In other words, the procedure has no provision for testing whether the items in the two tests have adequate fit to the Rasch model when combined to form one test. It would appear that this procedure might allow infringements to the Rasch model to go undetected.

It should be noted that the common item difference procedure could be carried out with common persons instead of common items. In this case, the approach involves the computation of the mean differences of the estimates of the common persons in the two tests being equated.

Concurrent equating

In a concurrent equating procedure, the data from two tests that are to be equated are combined to form one data set. The calibration of the two tests is then done simultaneously.

With the concurrent equating procedure, there is direct interaction in the data of the tests being equated. Hence, items that behave differently from the other items in the combined data can be identified and removed from the analysis. In addition, it is suspected that the concurrent procedure automatically solves the problem identified by Linacre and Wright (1989) and DeMars (2001 & 2002), which is associated with the differences in the distribution of the outcome variable in the different groups of students taking the tests. Linacre and Wright (1989) and DeMars (2001 & 2002) argue that when the different ability distributions of the students taking the tests are not taken into account, this could distort the resulting scale. However, this may not be the case when a concurrent procedure is employed because the data are analyzed as one, simultaneously. Nevertheless, there is need for a further study to investigate aspects of this problem.

Based on a concurrent equating procedure, it is not possible to estimate the errors of equating. Nevertheless, several research studies have shown that the concurrent method provides more consistent and stronger measures of the two sets of items and persons being equated (Kenyon and Stansfield, 1992; Morrison and Fitzpatrick, 1992; Baker and Al-Karni, 1993; Shen, 1993; and Mohandas 1996). For example, a study carried out by Kenyon and Stansfield (1992) to compare Rasch model vertical methods, demonstrated that concurrent equating has a beneficial effect on the calibration of the common items. As another example, a study carried out by Mohandas (1996) employed the Rasch model to equate five test forms using both concurrent and anchor item equating and found that the former technique yielded more consistent results.

In this study, the concurrent procedure is employed to link the Grade 3 to the Grade 5 tests on the same testing occasions. The concurrent procedure is also used to equate the combined Grades 3 and 5 data for 1996 to 2000. However, the common item difference procedure is used to link the 1996 to 2000 scale to the 1995 scale so as to obtain a common scale running from 1995 to 2000. In all the analyses, the equating of the numeracy tests is done separately from the equating of the literacy tests. The steps undertaken to equate the tests in this study are provided in more detail in Chapter 6.

Problems of equating tests using the Rasch model

Several studies have been carried out to investigate the appropriateness of using the Rasch model in equating. These studies have come up with apparently contradictory findings.

Some studies have provided evidence to oppose the use of the Rasch model especially in vertical equating. For example, Slinde and Linn (1978) explored the adequacy of the Rasch model for the problem of vertical equating. They concluded that despite the promising use of the model, empirical results raise questions about the adequacy of the Rasch model. They recommended the use of latent trait models with more parameters in vertical equating.

In another study, Slinde and Linn (1979) used the Rasch model to equate reading comprehension tests of widely different difficulty for three groups of fifth grade students of widely different performance levels. They concluded that under these extreme circumstances, the Rasch model equating was unsatisfactory. However,

Gustafsson (1979) used computer generated data to show that Slinde and Linn's criticism of the usefulness of the Rasch model for equating might have been the result of an artifact produced by the manner in which the samples were chosen in their study.

Loyd and Hoover (1980) used the Rasch model to equate three levels of a mathematics computation test. Sixth, seventh, and eighth grade students were administered different levels of the test. There was lack of consistency between equatings, which suggested that the Rasch model did not produce a satisfactory vertical equating of the computation test.

Holmes (1982) created two tests from a standardized reading achievement test and vertically equated them using a sample of third and fourth grade students. From the differences in performance estimates for the same student, Holmes concluded that the Rasch model did not provide a satisfactory means of vertical equating.

On the other hand, there are studies that have provided evidence to support the use of the Rasch model in vertical equating. For example, Schratz (1984) compared the results of vertical equating using the Rasch model with those obtained using traditional methods. She concluded that the Rasch model compared well to the traditional methods. She noted that the findings had encouraging implications for computerized adaptive testing and customized test development and scoring.

O'Brien and Tohn (1984) carried out a study to investigate the application of the Rasch model equating and equipercentile equating in vertical equating. The study was conducted to determine whether, based upon Rasch vertical equating, a local school district should administer out-of-level tests (tests not for an actual grade level) to exceptionally able students. A comparison was made between the school district's equating results and those of the test publisher's vertically scaled scores based on equipercentile equating. The results indicated the publisher's vertical scale was comparable to the scale estimated from the local school district through the use of Rasch equating.

Sontag (1984) found that the one-parameter model yielded more stable results than the two- and the three-parameter models in vertical scaling of the data collected in the IEA Six Subject Study across the 10-year-old, the 14-year-old, and terminal secondary school levels in the areas of science, reading comprehension, and word knowledge.

A study by Shen (1993) provided strong support for the use of the Rasch model One-Step (concurrent) equating in vertical equating. The Rasch model measurement program BIGSTEPS was used to calibrate simultaneously three parts of the National Board of Osteopathic Medical Examiners' (NBOME) examination. The data consisted of 2,814 items and 5,168 persons, and despite the large amount of missing data, the program converged smoothly. The study found that the distributions of person performance were not affected by the equating. Shen observed that Rasch measurement provided good person ability estimates on the whole examination, and consistent difficulty estimates for items. Therefore, Shen concluded that One-Step vertical equating using the Rasch model was a valid, efficient, and accurate way to construct a measure for longitudinal medical achievement studies.

Equating studies indicate that the controversy over the appropriateness of vertical equating is not entirely restricted to the use of the Rasch model. Several studies have indicated that vertical equating may not be appropriate regardless of which IRT models or CTT methods are used in equating (Slinde and Linn, 1977; Reckase, 1981; Gialluca, 1984; Skaggs and Robert, 1988; Glowacki, 1991; Smith and Kramer, 1992; Wright and Dorans, 1993).

It appears that the main problem in vertical equating could be due to the possibility that the students at different grade levels might not respond to the two tests on the same dimension (Petersen et al., 1989). This problem stems from the fact that on a multidimensional test, two people could receive the same score for different reasons. However, Petersen et al. argue that as long as the scores on two multidimensional tests satisfy the other conditions of equating (such as equity, population-invariance, and symmetry) "it would be a matter of indifference to each examinee which form of test she or he took" (p. 243).

In the sections that follow, attention is focused on what past studies have said or found regarding school effectiveness issues that are of interest in this study: namely, (a) issues for research into school effectiveness, (b) school effect indices, and (c) multilevel modelling in school effectiveness research (SER). However, a short section in which models in student learning and school effectiveness are discussed precedes reviews on these issues.

Models in student learning and school effectiveness

Creemers et al. (2002; p.283) argue that, in respect to educational effectiveness research, a model is useful because it can "explain differences in student learning results by specifying the relationship between the components in the model and student outcomes". The aim of this section is to provide a conceptual base for the student achievement and school effectiveness models employed in this study. Details of the specific models that are examined in this study are presented in Chapter 5.

It should be borne in mind that the main purpose of this study is not to develop new methods of measuring school effects or to dispute the existing methods, but rather to investigate using existing methods for how these effects, for schools in South Australia, can be measured based on students' scores from the BSTP. In other words, the purpose of this study is not to develop entirely new methods of explaining the differences in student achievement but rather to develop models based on the knowledge gained from theory and previous research.

Educational researchers have advanced many different models of student learning in the last four decades especially after the development of the first model of student learning by Carroll in 1963 in which learning rate is considered as a function of five elements: aptitude, ability, perseverance, opportunity and quality of instruction. Specifically, Carroll's (1963) model states that student achievement (that is, success in learning) is a function of time actually spent divided by the time needed by a student. In this model, Carroll argues that both the time needed and the actual time spent are influenced by factors at the student-level such as the learner's ability and factors at the class-level (or group-level) such as quality of instruction. An empirical study carried out by Carroll and Spearritt (1967) generally confirmed the relationships hypothesized in Carroll's model.

Carroll's model has served as the foundation for the development of other models of student learning involving home and school environments (Bloom, 1976), time (Harnischfeger and Wiley, 1976), instruction (Harnischfeger and Wiley, 1978), perseverance (Keller, 1983), education productivity (Walberg, 1991), and student aptitude (Reynolds and Walberg, 1991). Indeed, Carroll's model has served as the basis for the development of models of student learning in specific school subjects such as mathematics and science (Keeves, 1975), science (Keeves, 1992b; Kotte, 1992) and reading (Lietz, 1996).

Creemers et al. (2002) attribute the success of Carroll's model in influencing further research in student learning to the fact that the model is directed towards what happens in schools unlike other models that are concerned with the learner and internal processes of learning models (e.g. Gage, 1963). The fact that Carroll's model is directed towards what happens in schools coupled with the fact that the model offers a set of relevant factors at the student and the group-level have made the model very attractive to school effectiveness researchers. Indeed, Creemers et al. (2002) have noted that virtually all multilevel school effectiveness models refer to Carroll's (1963) model of student learning.

Nevertheless, Creemers et al. (2002) have pointed out some drawbacks of Carroll's model in school effectiveness research, the main one being the failure of the model to pay much attention to the definition of the factors at the group-level. Because of this drawback, some researchers have extended Carroll's (1963) model to put more emphasis both on the school and class levels (e.g. Stringfield and Slavin, 1992), some have placed more emphasis at the school-level (e.g. Willms and Raudenbush, 1989; Scheerens, 1992), while others are more interested in what happens at the class-level (e.g. Creemers 1994). In addition, some researchers have extended Carroll's model to include factors at levels above that of the school such as district, state and federal levels (e.g. Stringfield and Slavin, 1992; Creemers, 1994; Creemers et al., 2002).

Creemers' (1994) model of educational effectiveness is an extension of Carroll's model that has attracted substantial interest from researchers into school effectiveness. The Creemers' model draws attention to what happens at the class-level, and connects this to what happens at the school-level as well as the interaction between the class and school levels. This model also focuses on what happens at levels above the school-level. Creemers (1994) refers to levels above the school-level as 'context level'. He argues that context level factors that influence student learning could include national policies that focus on the effectiveness of education, teacher training and funding of schools based on outcomes.

For the current study, Creemers' (1994) model can not be employed because BSTP data lack information at the class-level.

The Willms and Raudenbush (1989) and Raudenbush and Willms (1995) model of school effectiveness is an extension of Carroll's model, which draws special attention to what happens at the school-level. At the school-level, this model differentiates between so-called 'school context' variables and 'school policy and practice' variables. Willms and his colleague say that school context variables consist of aspects of school environment that are thought to influence student achievement and are not under the direct control of the school staff. Such aspects may include the average student characteristic variables, such as the average school prior achievement and the average school socioeconomic status. On the other hand, they say that school policy and practice variables consist of aspects of the school that are thought to influence student achievement and are under direct control of the school staff, such as school leadership, curricular content, instructional quality, and resource use.

There are at least two reasons why the above model by Willms and Raudenbush is interesting. First, this model separates the effects of school context from the effects of school policy and practice and, therefore, it allows the researcher to identify schools whose policies and practices appear to promote best student achievement after allowance has been made for student background and school context factors (Pituch, 1999). Second, the statistical method employed in this model allows the researcher to estimate the effects of individual school policies and practices without necessarily obtaining the measures of these policies and practices. Raudenbush and Willms (1995) refer to this statistical method of estimating the effect of individual school

policies and practices as the 'subtraction' approach. The alternative statistical method is referred to as the 'addition' approach, that requires the specific policies and practices to be measured and the resulting variables to be entered into the multilevel regression analysis employed to estimate school effects. The subtraction method is very useful because in real life situations these policies and practices are problematic to measure (Willms, 1992; Raudenbush and Willms, 1995).

In the BSTP, no data are collected on school policies and practices and, consequently, there is no information regarding individual school policies and practices in the data available for the current study. Because based on the above model by Willms and Raudenbush it is possible to estimate the effect of individual school policies and practices by using the subtraction approach, this model was chosen for study. More details regarding this model and its use in this study are presented below (2.6) and in Chapter 4 (4.2).

Issues for research in school effectiveness

The question of whether schools influence their students' academic achievement has interested many researchers. The study by Coleman et al. (1966) is among the earliest influential works on the role of the school in student achievement, followed by Peaker (1967) and Jencks et al. (1972). Like other early studies of the 1960s and 1970s, these studies were based around conventional multiple regression techniques and concentrated mainly on relationships among student-level variables. A major shortcoming of the these early studies was a failure to model for the way in which students were allocated to schools, which meant that the resulting statistical inferences were biased and, moreover, the statistical model could not untangle the influence of the school as such (Goldstein, 1997).

By the late 1980s, researchers were using multilevel analysis techniques to study the role of schools on student achievement (e.g. Mortimore et al., 1988), and developing a new literature, a new language and a new discipline of 'school effectiveness' (e.g. Aitkin and Longford, 1986; Goldstein, 1987; Raudenbush, 1988). In the early 1990s, school effectiveness research (SER) was fast becoming an applied discipline with researchers (especially in Britain) becoming involved in the production and use of value added scores, usually measures of the contribution of the school to the increase in student achievement. In the 1990s, researchers into school effectiveness became interested in examining changes in school performance over time, thus looking at school improvement from the perspective of school effectiveness (e.g. Teddlie and Stringfield, 1993; Gray et al., 1995, 1996 & 1999).

In the pioneering studies of school effectiveness, inquiries seem to have been guided by the question of whether or not schools influence their students' academic achievement (e.g. Coleman et al., 1966; Jencks et al., 1972). Soon substantial evidence became available that schools do actually influence their students' academic achievement (e.g. Mortimore et al., 1988; Scheerens, 1992; Teddlie and Stringfield, 1993) and researchers started to focus their attention on other questions and issues.

However, it should be noted that SER has generally remained focused on two main objectives. The first objective is to identify unusual schools (that is, extremely effective or extremely ineffective schools) and this identification usually serves as a first step in qualitative research, which usually includes on-site visits to examine these schools more carefully. The second objective is to identify school characteristics that lead to differential student outcomes. Within these two research objectives, there are a number of issues or research questions involved. Brief discussions focusing on past SER that have addressed issues that are of interest in the current study are provided

next. A comprehensive treatment of school effectiveness research can be found in a recent volume edited by Teddlie and Reynolds (2001a). Books by Slee et al. (1998) and Thrupp (1999) contain criticisms of SER (mostly political and mainly from Britain) while articles by Teddlie and Reynolds (2001b), and Reynolds and Teddlie (2001) have countered the criticisms contained in these two books. Two recent books by Saunders (1998 & 1999) provide overviews and critical reviews of SER (mainly in Britain). Articles by Coe and Fitz-Gibbon (1998), Thrupp (2001), and Scheerens et al. (2001) provide some insights into the shortcomings of SER.

Variance and magnitude of school effect

Researchers are interested in the question of how much difference schools make to the variance in student achievement because variance between schools has practical implications for parents, administrators, policy makers and others who are concerned with school learning. For parents, if the variance between schools is zero, there would be no consequences for the expected achievement of a child when choosing among a set of schools; whereas, if the variance is large, such choices would be of crucial importance (Raudenbush and Willms, 1995). For policy makers and administrators, the magnitude of the variation between schools is important because it is an indicator "of the extent of inequality produced by the schooling system" (Raudenbush and Willms, 1995; p. 315).

A number of studies have indicated that schools make substantial contributions to the variance in student achievement. For example, Gray et al. (1995) found that around 19 to 20 per cent of the variation in performance of students in the 1990, 1991 and 1992 GCSE examination could be attributed to the differences between secondary schools in England.

There is substantial evidence that the proportion of variance in students' achievement associated with differences between schools is relatively larger in some countries compared to the proportion of variance in other countries. For instance, a multilevel analysis of data from South Africa (a developing country), which were collected as part of TIMSS-R, revealed that 55 per cent of variance in mathematics score lay at the school level (Howie, 2002). A similar analysis of data from Indonesia (also a developing country) collected as part of TIMSS revealed that 44 per cent of variance in mathematics score was at the school level (Mohandas, 1999). Willms and Somers (2001) reported similar findings in their recent work with Grades 3 and 4 pupils from 13 Latin American countries. Using Australian data collected as part of PISA, Lokan et al. (2001) found that 17 per cent of variance in 15 years olds' reading scores could be attributed to between school differences, which is a relatively small proportion of variance compared to the OECD average (36 per cent). Generally, however, it would appear that in developed countries there is not much difference between schools but in developing countries the difference between schools could be large.

There are also indications that, within the same country, the variance between schools appears to be different for different school subjects (see Thomas et al., 1997; p.186; Willms and Somers, 2001; p.419).

Raudenbush and Willms (1995; pp.316-7) cautioned against assuming that the amount of variation between schools puts an upper limit on the variance of school effects. They argued that the variance attributed to school effects could be larger than the overall variation between schools for a number of reasons, one of the reasons being that school effects can influence within school variance by interacting with student background. In addition, it is generally agreed that the proportion of variance explained by school-level variables is a poor guide to the real influence of schools to

the increase in student achievement (e.g. Rutter, 1983a; Bosker and Scheerens, 1989; Scheerens, 1992; Sammons et al., 1996; Willms, 1992; Teddlie et al., 2001). Consequently, a number of alternatives have been proposed for expressing the magnitude of school effects (see Teddlie et al., 2001; pp.102-4). Of considerable popularity is an approach proposed by Willms (1992; p.43) that involves expressing the magnitude of school effect as a fraction of the standard deviation of the outcome measure to yield what he calls an 'effect size'. Based on the approach proposed by Willms, the sign of an effect size can be positive or negative, meaning that an individual school is either more or less effective for an individual student compared to other schools. For example, an effect size of 0.1 means that the school is more effective by 10 per cent of a standard deviation (on the original outcome scale) compared to other schools included in the analysis. Bosker and Witziers (1996) and also Brandsma and Doolaard (1999) argue that effect sizes could be more relevant if expressed in terms of years of life that a student spends at school.

Differential school effects

Some research findings have indicated that schools could be differentially effective in that they appeared to promote with different effectiveness the academic achievement of different groups of students, divided by such characteristics as prior achievement, socioeconomic status and ethnicity (e.g. Nuttall et al., 1989; Willms and Chen, 1989; Young and Fraser, 1993; Thanassoulis, 1996; Pituch, 1999). Differential school effects upon students with different characteristics within schools relates to the issue of consistency of school effects across subgroups of students and should not be confused with the issue of contextual effects. Teddlie et al. (2001) note that:

> Contextual effects are related to the overall composition of the student body (e.g. the percentage of high ability or of high SES students in a given year group or in the school's intake as a whole) and can be identified by between school analyses across a sample of schools. (Teddlie et al., 2001; p.127)

They continue to note that research studies in secondary schools in the United Kingdom have suggested that "contextual effects related to concentrations of low SES, low ability and ethnic minority can be important" (p.127). The current study is mainly interested in the issue of differential school effects. However, issues related to contextual effects are also considered, especially when interpreting cross-level interaction effects in this study.

For prior achievement, evidence of differential school effectiveness is available in the secondary sector (Willms and Raudenbush, 1989; Nuttall et al., 1989; Nuttall, 1990), as well as in the primary sector (Sammons et al., 1993). However, some studies have reported a lack of conclusive evidence for the existence of differential school effectiveness based on prior achievement levels of the students at the secondary sector (Jesson and Gray, 1991), and also in the primary school sector (Brandsma and Knuver, 1989). More recently, a secondary school study by Harker and Nash (1996) in New Zealand reported some evidence of differential school effectiveness for students of different prior achievement levels in science and English but not in mathematics.

For gender, ethnicity and socioeconomic status, some studies have reported evidence to support the existence of differential school effects at the secondary school level (Nuttall et al., 1989; Pituch, 1999) but lack of such evidence was reported by studies that examined these issues at the primary school level (Sammons et al., 1993). In addition, some secondary school studies have reported lack of substantial evidence to support the existence of differential school effects related to student gender (Willms

and Raudenbush, 1989; Harker and Nash, 1996) or ethnic background (Harker and Nash, 1996) or socioeconomic status (Willms and Raudenbush, 1989; Harker and Nash, 1996).

The study by Pituch (1999) used a subset of data from the United States National Longitudinal Studies Program (NLSP) to illustrate that the ranking of 96 schools based on their contribution to the increase in mathematics achievements of Grade 10 students changed with different levels of student socioeconomic status considered. He reports that around 95 per cent of the 96 schools "change between 1 to 10 ranks, and 5% of the schools change between 11 to 20 places when students SES scores change from one standard deviation to the mean, with a median change of three positions" (Pituch, 1999; p.199). However, it should be noted that schools included in this illustration were purposely selected to demonstrate the consequences of ignoring differential effectiveness in ranking of schools and, therefore, it is unlikely that such substantial changes in ranks would occur in non-manipulated data situations.

Consistency of school effects across outcome measures

The issue of consistency of school effects across outcome measures relates to correlations among school effects across outcome measures rather than the relative magnitudes of the school effects across outcome measures. The current study is interested in correlations among school effects across two outcome measures, that is, numeracy and literacy.

A number of researchers have found evidence that school effectiveness is outcome specific: that is, some schools may perform relatively better in one outcome (e.g. numeracy) and relatively poorer in another outcome (e.g. literacy). Consequently, some researchers have cautioned against the idea of attempting to capture the effectiveness of a school with a single summary index (e.g. Nuttall et al., 1989; Willms and Raudenbush, 1989; Fitz-Gibbon, 1991; Raudenbush and Willms, 1995; Thomas et al., 1997; Coe and Fitz-Gibbon, 1998; Pituch, 1999). Generally, it appears that correlations across different subject areas at the primary school level are higher than at secondary school level.

A study of Scottish secondary schools by Cuttance (1987) reported correlations between school effects on an overall measurement and effects on English and arithmetic achievement of 0.47 and 0.47 respectively. Another Scottish study that employed data consisting of two cohorts of students who completed their secondary school in 1980 and 1984 found that the correlation between school effects (adjusted for student intake) on English and arithmetic were 0.46 for 1980 and 0.73 for 1984 (Willms and Raudenbush, 1989). More recently, Thomas and Mortimore (1996) reported a correlation of 0.46 between value added measures of GCSE English and mathematics, and Thomas et al. (1997) reported a correlation of 0.35 between school effects on science and mathematics and a correlation of 0.72 between school effects on English and English literature. From the study by Thomas and Mortimore, it appears that there is higher degree of consistency between school effects based on outcome measures that are very similar compared to the consistency of school effects based on outcome measures that are dissimilar.

At the primary school level, studies in the United States conducted by Mandeville and Anderson (1987) and Mandeville (1988) reported correlations between school effects for mathematics and reading in the 0.60 to 0.70 range while a British study by Sammons et al. (1993) reported a correlation of 0.61 between school effects on mathematics and writing. Similarly, Bosker and Scheerens (1989) using data collected from elementary schools in the Netherlands reported a strong positive correlation of

0.72. In Australia, so far there are very few studies that have adequately explored the issue of consistency of school effects across outcomes.

Stability of school effects over time

The issue of stability of school effects over time relates to the consistency of school effects within the same subject over time. There are two possible questions associated with this issue. There is the question of consistency of school effects across grades within the same subject area, and there is the question of consistency of school effects within the same grade and same subject areas over time. The current study is interested in the second question; namely, 'Are schools that are effective for one cohort of students in numeracy (or literacy) also effective in numeracy (or literacy) for subsequent cohorts of students?' Thus, in order to address the question of stability of school effects within the same grade over time, it requires that data should be collected on more than one cohort of students. Consequently, there are very few studies that have addressed the question of stability of school effects within the same grade over time, simply because of the time required for the collection of data from more than one cohort of students. Nevertheless, in large scale testing programs (such as the South Australian BSTP), the data collected on different testing occasions could be used to examine the stability of school effects over time.

Past research studies have not yielded clear cut evidence on how stable school effects are within the same grade across time, but suggest that schools that are effective for one cohort of students are generally also effective for other cohorts of students (e.g. Nuttall et al., 1989; Willms and Raudenbush, 1989; Sime and Gray, 1991; Luyten, 1994a; Gray et al., 1995 & 1996; Thomas et al., 1997). Results from past studies also suggest that school effects based on an overall measure of academic outcome are more stable than those based on specific academic subjects (e.g. Willms and Raudenbush, 1989; Gray et al., 1995; Thomas et al., 1997). However, most studies on stability of school effects over time have focused on performance of secondary schools. In addition, of these secondary school studies, very few have examined the stability of school effects for a range of academic and non-academic outcomes.

A Scottish secondary school study by Raudenbush (1989) reported a correlation of 0.87 between school effects on an overall achievement score, and therefore concluded that school effects on overall achievement are fairly stable. Raudenbush (1989) and Willms and Raudenbush (1989) argued that research on stability of school effects ought to differentiate between instability due to true changes in school performance and instability due to measurement and sampling errors. They drew attention to the importance of adopting a longitudinal model for estimating school effects and examining their stability over time. This idea of using a series of cohorts in the estimation of school effects has also been recommended by a few other researchers (e.g. Nuttall et al., 1989; Fitz-Gibbon, 1991; Teddlie et al., 2001).

Similar to the Scottish study by Raudenbush, a secondary school study by Gray et al. (1995) in the United Kingdom found that school effects based on 1990, 1991 and 1992 GCSE total scores were considerably stable from year-to-year, with a range of correlations from 0.81 to 0.96. However, Gray and his colleagues also found evidence of changes in school effectiveness over time, a point they seem to confirm in a subsequent study where they used data from five testing occasions (Gray et al., 1996). Another United Kingdom study by Thomas et al. (1997) examined the stability of school effects across three GCSE cohorts (1990-1992) on seven outcome measures and reported that, within the same outcome measure, the relationship between school effects varied considerably as "illustrated by a range of correlations from 0.38 to 0.92" (p.190). Consequently, Thomas et al. (1997) concluded that there was a substantial

degree of change over time in some schools and recommended looking at the results of school effects over several years.

As noted above, only a few studies have examined the issue of stability of school effects within the same grade over time at the primary school level. Nevertheless, results from the few past studies that have examined this issue indicate that the correlations are lower compared to those observed in the secondary sector. For example, work by Mandeville and Anderson (1987; p.212) in the United States found that the correlations between school effects within the same grade for the subject area (reading and mathematics) were all "discouragingly small [less than 0.20], with the majority not achieving statistical significance at the .05 level". However, a recent study by Crone et al. (1994) in the United States reported much higher correlations (between 0.49 and 0.78) compared to the correlations reported by Mandeville and Anderson, suggesting existence of substantial stability in school effects over time at the primary school level.

In spite of the recommendation by Willms and Raudenbush (1989), most studies having found evidence of instability in school effects over time have not proceeded to separate the true changes in performance of the school from the sampling and measurement errors. Indeed, very few studies have attempted to examine changes in school performance over time. Gray et al. (1995) argue that most studies do not examine these changes because they either consider the changes to be small or the studies are more focused on replicating their findings and, therefore, see instability across years as a threat to their findings. However, Gray and his colleagues argue convincingly that instability is essential for study of change. They proceed to outline a number of factors that must be brought together in order to conduct a satisfactory study of changes in school performance over time. In their view these factors include:

- measures of outcomes and prior attainment on individual pupils;
- data on a minimum of three cohorts and preferably more;
- a multi-level statistical analysis;
- an orientation towards examining data for systematic changes in schools' performance over time. (Gray et al., 1995; p.100)

Based on the above criterion, the current study is adequately positioned to examine the issues of changes in school performance over time for the primary schools in South Australia on two outcome measures, numeracy and literacy. Furthermore, unlike the few previous studies that have attempted to tackle the problem (e.g. Gray et al., 1995 & 1996), in the current study the outcome and prior achievement can be measured on the same scale and, therefore, this study should provide a clearer picture of the stability and changes in school performance over time.

School effect indices

Within a value-added framework, school effect indices or indicators (SEIs) have been defined by William et al. (2000; p.1) as "the differences between the school's actual mean performance and the school's expected performance based on the achievement of other schools with similar levels of student and school characteristics". Others have defined SEIs as statistics that are collected at regular intervals to track an education system (e.g. Fitz-Gibbon, 1990; Fitz-Gibbon and Kochan, 2001). Accordingly, for the purposes of the current study, SEIs are simply defined as statistics that describe (or provide a summary) of the performance of a school in a specific outcome measure. Fitz-Gibbon and Kochan (2001; pp.257-282) have provided a thorough treatment on definitions of school effectiveness indicators, comprehensive descriptions of the

criteria for selecting the indicators, categories of indicators and examples of official indicator systems from several countries.

The next two sub-sections focus on some SEIs that have been identified by past studies and what the past studies have identified as the common problems associated with the estimation of SEIs.

Types of school effect indices

Hill (1996), and also Hill and Rowe (1996), have identified three types of value added indicators that are commonly used to describe school effectiveness, namely: (a) unpredicted achievement, (b) learning gain, and (c) net progress.

The 'unpredicted achievement' indicator describes the achievement level adjusted for family background and ability. The 'learning gain' indicator describes the achievement level adjusted for initial achievement level. The 'net progress' indicator describes the achievement level adjusted for family background, ability and initial achievement. Hill (1996) and Hill and Rowe (1996) predict that these three school effectiveness indicators are fairly highly correlated but caution that they relate to different aspects of educational effectiveness. They recommend that where possible it would be more appropriate to report school effectiveness based on more than one kind of these indicators. Nevertheless, they emphasize that the validity of any value added indicator depends upon the extent to which adjustments have been made for all relevant intake characteristics and that each of these have been measured reliably.

Willms and Raudenbush (1989) and Raudenbush and Willms (1995) have identified two kinds of school value-added indicators that have attracted many school effectiveness researchers. The two indicators are namely: 'Type A' and 'Type B' effects. A Type A effect indicator describes the contribution of a given school to the increase in student achievement after controlling for all student-level factors influencing student achievement such as entry achievement level, student family background, and so on. Meyer (1996 & 1997) refers to the Type A measure as a total school performance indicator because it accommodates all (internal and external) school-level factors that influence the increase in student achievement. Ideally, it is possible to use the Type A effect to report value added indicators for different types of students in a particular school (Meyer, 1996; p.201). Thus, parents choosing a school for their children would be interested in a Type A effect indicator (Raudenbush and Willms, 1995; Harker and Nash, 1996).

On the other hand, the Type B effect indicator reflects the contribution of a given school to the increase in student achievement after controlling for student-level factors and external[3] school-level factors that influence the increase in student performance. Meyer (1996 & 1997) refers to the Type B measure as an intrinsic school performance indicator because it accommodates only the internal[4] school-level factors that influence the increase in student achievement. Thus, the Type B effect serves as a good indicator for the purpose of holding schools accountable for their performance (Harker and Nash, 1996).

In addition, Willms and Raudenbush (1989) have argued that where data are available on more than one cohort of students within the same year level, each of the two types

[3] Observable school characteristics that can be considered external to a particular school, principally neighborhood and community characteristics and aggregated student characteristics (Meyer, 1996; p. 202).

[4] Observable school characteristics that can be considered internal to a particular school, principally school policies and inputs (Meyer, 1996; p. 202).

of indicator (that is, either Type A or Type B) could be split into a 'stable' component and an 'unstable' or change component. They argued this splitting would provide information regarding the overall performance of the school and change (improvement or deterioration) in school performance over the study period. Gray et al. (1995, 1996 & 1999) have also argued along this line.

Indeed, there are many indicators that have been used to describe school effectiveness (see Fitz-Gibbon and Kochan, 2001; pp.270-282). Meyer (1996; p.178) pointed out three attributes that a performance indicator must possess in order for the indicator to be acceptable and valid. Meyer specifically identifies the three attributes as (a) outcome validity, (b) noncorruptability, and (c) valid measurement of school performance. In addition, McPherson (1996; p.2) proposed that any assessment of school effects on student progress could be seen as valid so long as it is based on what he calls an "explicit theory of good standing". McPherson contends that any such theory should take into account that (a) schooling is longitudinal, (b) schooling is multilevel, and (c) there are many factors involved in student achievement.

Problems in estimation of indices of school effects

Several concerns have been raised regarding the estimation of school effects. Outstanding among these concerns are (a) issues to do with bias, and (b) variables to include (or exclude) in the analysis when estimating the indices.

Raudenbush and Willms (1995) and Kennedy and Mandeville (2001) have described the problems associated with the estimations of Type A and Type B school effects. They note that for purposes of drawing causal inferences about school effects, students should be randomly assigned to schools and schools should be randomly assigned to context and process conditions. However, they note that these random conditions are not possible in real life situations, and therefore, they conclude that it is difficult to estimate either type of school effects without bias, but it is relatively less challenging to estimate Type A effects without bias than to estimate Type B effects without bias.

In the absence of the random allocation of students to schools, the estimation of Type A effects could proceed without bias if relevant data on student background characteristics were measured accurately and included in the model. Likewise, the estimation of Type B effects could proceed without bias if the relevant data on student characteristics, school context and policy or practice were measured accurately and included in the model. However, the estimation of Type B school effects without bias would be further complicated in that it would require that school context and policy were orthogonal. Because school context and policy are in some cases related, Raudenbush and Willms conclude that Type B school effects tend to be estimated with some bias (see Raudenbush and Willms, 1995; pp. 318-319, for a more extented discussion of this issue).

Another concern associated with the estimation of school effects relates to the variables to include in the model. Research findings based on conventional regression modelling techniques suggest that inclusion of different sets of predictor variables, even those that may contribute a small amount to variance explained, may lead to different estimates of school effects and could yield different sets of rankings (Douglas, K., 1988). However, there are also indications that as the number of variables in the analysis increase, the reliability of the estimates decreases (Coe and Fitz-Gibbon, 1998).

Nevertheless, it is generally agreed that school effects that are unadjusted for the intake characteristics of the students are biased because a good school with a disadvantaged student intake could never perform as well as a mediocre school with

the so-called 'head start' of a more able population (Coe and Fitz-Gibbon, 1998). Ideally, for estimation of Type A school effects to proceed without bias, it is necessary that every relevant aspect of an individual student should be measured accurately and included in the model. However, in reality, it is very difficult to measure accurately all relevant aspects of individual students and, therefore, "any measure of value added which we calculate may be thought of as an attempt to measure 'pure' value added that is biased towards unadjusted (raw) performance" (Coe and Fitz-Gibbon, 1998; p.425). Nevertheless, Coe and Fitz-Gibbon (1998) add that studies which, for example, have no measure of SES but adjust only for prior achievement may be closer to the measuring of the actual effects of schooling than studies which have no measure of prior achievement and adjust only for SES.

Despite what has been said above, research has shown that measures of prior achievement and SES background of the student are in most cases sufficient control because they capture the bulk of the within and between schools variations (Willms, 1992). Nevertheless, there is clearly a need for studies in this area to use multilevel modelling techniques in order to examine the impact on the ranking order of the schools when some predictors are omitted in the analysis.

Multilevel modelling

In this sub-section, a review on issues of multilevel modelling (MM) techniques that are of interest in this study is provided. The first part of this section focuses on what past studies have said regarding the importance of MM techniques in SER while the second part focuses on the problems that past studies have associated with MM techniques.

Importance of multilevel modelling

Kennedy and Mandeville (2001; p.190) have noted that conventional techniques used in SER "either required that the investigators ignore the multilevel nature of school data or to incorporate multiple levels in ways that were technically questionable". They have identified a number of conventional techniques (e.g. single-level regression, contextual analysis and slope as outcome) used in earlier SER and outlined the limitations associated with these techniques; namely, aggregation bias and misleading parameter estimates. Cheung et al. (1990; pp. 215-319) have comprehensively addressed the issue of aggregation bias (or 'grouping effects' as they prefer to call it).

Unlike conventional techniques, multilevel technique "attempts to more realistically reflect the nested or hierarchical nature of data encountered in school effects studies" (Kennedy and Mandeville, 2001; p.191). The hierarchical nature of school effects occurs because of the characteristics of students being taught within classes, within schools, that are nested within districts and within provinces.

Consequently, most researchers agree that studies that employ multilevel models that represent the hierarchical nature of schooling are more appropriate in the estimation of school effects (e.g. Willms and Raudenbush, 1989; Teddlie et al., 2001). Indeed, the development of powerful computer packages such as HLM5 (Raudenbush et al., 2000), MLwiN (Browne et al., 2001) and VARCL (Longford, 1990) has lead to numerous school effectiveness studies that have employed MM techniques especially in recent years. Hox (1995) has reported comparisons of earlier versions of these three multilevel programs, and notes that the three programs yield similar results when the data sets involved are relatively large. Comparisons of earlier versions VARCL, MLwiN and other MM softwares have also been reported by Cheung et al. (1990) and

Kreft et al. (1990 & 1994). Generally, these computer programs are appealing to researchers because they enable statistical estimations of regression coefficients and also provide appropriate standard errors (Bryk et al., 1994 & 1996; Raudenbush et al., 2000), confidence intervals and significance tests.

Despite the merits associated with multilevel modelling techniques in the estimation of school effects, structural equation modelling (SEM) seems to be gaining widespread popularity among some educational researchers. Kennedy and Mandeville (2001) say that the advantages of SEM technique include:

- explicit tests of factor structures and their invariance over groups;
- explicit incorporation of unique and correlated measurement errors in analyses;
- comparisons of means of latent construct, not observed variables contaminated by measurement errors, and
- provision for a systematic approach to hypothesis testing. (Kennedy and Mandeville, 2001; p.202)

However, Kennedy and Mandeville (2001) note that most SEM applications assume that there is no hierarchical ordering of observations in the population studied and, consequently, they argue that this assumption could be questionable especially in school effectiveness research. Nevertheless, they note that recent developments in SEM have begun to address multilevel issues. Thus, SEM techniques could gain popularity among school effectiveness researchers in the near future.

Problems in multilevel modelling

Some questions have been raised regarding the validity of the conclusions reached using multilevel modelling (MM) techniques. Kreft et al. (1995) report that the common practice of centring the predictors in multilevel modelling yields results that may differ from the raw score predictors. Consequently, Kreft et al. (1995; p.15) recommends that researchers should have conceptual reasons for using raw scores or centred ones in a given research analysis since general reasons or rules for centring can not be given. Others have also given this same recommendation (see Bryk and Raudenbush, 1992; Raudenbush et al., 2000; Kennedy and Mandeville, 2001).

Apart from centring, there are a number of other issues that have been raised regarding the appropriateness of MM techniques. For example, a simulation study by Marsh (1998) evaluated path analysis and growth modelling approaches to multilevel change in relation to ubiquitous regression to the mean problem. The study by Marsh provided support for the validity of MM approach to change, but questioned the appropriateness of the interpretation based on multilevel growth modelling approach. Marsh simulated data with students assigned to schools on the basis of pretest (T1) scores so that there were moderately large initial differences in school-average achievement but individual growth did not vary from school to school. The results of the multilevel path analysis approach were consistent with how the data were constructed. Student growth in achievement did not vary with school-average achievement. On the other hand, the results of the multilevel growth model (multilevel repeat measures) approach "implied that there was substantial school to school variation in achievement growth over time and this school variation was completely explained by the pretest (T1) school-average achievement" (Marsh, 1998; p.10). However, Marsh cautions on the generalizability of the findings because the simulated conditions were highly unlikely to occur in actual practice.

Some critics argue that multilevel modelling techniques have no advantage over traditional regression techniques. For example, simulation studies carried out by Kreft

(1996) indicated that regression parameters obtained using multilevel analysis were close to those that were obtained using traditional regression techniques. However, Kreft's simulation studies established that the multilevel techniques were superior in the estimation of the standard errors of the parameters compared to the traditional regression techniques. A number of other studies have also come up with similar findings (e.g. De Leeuw and Kreft, 1995a & b; Rogosa and Saner, 1995; Trower and Vincent, 1995; Kennedy et al., 1993).

In addition, regarding the ranking of schools, a number of researchers have reported that the results obtained using less complex techniques do not differ markedly from those obtained using MM techniques (e.g. Webster et al., 1995; Fitz-Gibbon, 1996; William et al., 2000). Webster et al. (1995) examined the ranking order of schools obtained using the student-based regression modelling technique used by DISD[5] in Texas in the United States and the ranking order of the schools obtained using a number of two-level hierarchical linear modelling techniques. They concluded that the student-based model and the two-level hierarchical linear models produced very similar school ranks (r in excess of 0.90).

More recently, William et al. (2000), using data from three forms of a test administered to Grades 3 and 5 students in Maryland (USA), reported extremely strong correlations (r in excess of 0.90) between the school effects obtained using hierarchical linear models and school effects obtained using student-based regression models. However, on the basis of stability of the results across the three forms of the test, they recommended that the hierarchical modelling approach should be used for estimating school effects.

Another issue associated with the use of multilevel models in SER that has attracted some criticisms is referred to as 'shrinkage' of school effects values associated with individual schools. The idea of shrinkage is simply an adjustment of school effects to cater for sampling and measurement errors. Willms (1992; p.46) says that during this adjustment the estimates of school effects are said to be "shrunk" towards the mean outcome score for the entire sample and that the shrinkage of estimates for small schools is greater than the shrinkage for large schools. He argues that shrinkage presents "a more accurate picture of the variation between schools" regardless of whether representative samples or whether entire schools are included in the analysis (see Willms, 1992; p.42 for an extended discussion on this issue). However, others argue that shrinkage is a less justifiable adjustment when entire schools and classes are used (Fitz-Gibbon, 1996). Some argue that this adjustment could be frustrating to principals of excellent schools who might not wish to see the success of their students shrunken towards the mean (De Leeuw and Kreft, 1995a).

Another issue of concern associated with multilevel modelling in SER relates to the number of levels to include in the study. Generally, there are concerns that the results of the analysis could be misleading (or less informative) if important levels of a hierarchy are omitted (Teddlie and Stringfield, 1993). However, there are indications that as the number of levels increase, the stability of the results will decrease (Morris, 1995) and the more difficult it might be to comprehend the results (Tymms, 1994; De Leeuw and Kreft, 1995b; Willms and Kerckhoff, 1995; Baker et al., 1995) making it difficult for the results to have a practical impact (Bock and Wolfe, 1996; Teddlie et al., 2001). It is generally agreed that the nature of the data and the purpose of the analysis should guide the selection of the number of the levels to include. In addition, Kennedy and Mandeville (2001; p.198) argue that "the sample should be sufficiently large to permit simultaneous estimation at each level of hierarchy studied".

[5] Dallas Independent School District.

Nevertheless, there does not appear to be disagreement that multilevel models allow the exploration of the extent to which differences in achievement between students can be accounted by factors such as classroom characteristics, school characteristics, gender and other background characteristics of the students. Past studies have clearly demonstrated how MM techniques can be applied to explain why various aspects of the schools differ for the different kinds of students (Goldstein et al., 1993; Willms and Somers, 2001).

Issues in modelling for school effectiveness in this study

It has been mentioned above that the model proposed by Willms and Raudenbush (1989) and Raudenbush and Willms (1995) is chosen for estimation of school effects in this study. This model is chosen because it is directed at what happens at the school-level, which is the level of interest in the current study (see Chapter 1). In addition, the data available for the current study has no information at the class-level and, therefore, models of school effectiveness that include a class-level (e.g. Creemers, 1994) cannot be employed in this study. Moreover, the data available for this study do not have information to facilitate the construction of school policy and practice variables. And based on the model by Willms and Raudenbush (1989), it is possible to estimate the effect of individual school policies and practices without necessarily obtaining the measures of these variables.

Regardless of the appropriateness or inappropriateness of the model chosen for estimation of school effects, there are issues concerning some of student-level variables included in analyses that warrant some discussion. For the current study, variables that could be in question are (a) prior achievement (b) socioeconomic status, and (c) transience. In the sub-sections that follow, issues related to these three variables in the current study are discussed. The data available for this study are described in details in the next chapter together with the all the other variables involved in this study. The specific models used to estimate the school effects in this study are described in Chapter 5.

Prior achievement

Nearly all studies on factors influencing student achievement have shown that prior achievement is highly correlated with later student achievement, with students with high prior achievement scores achieving higher scores in subsequent achievement tests. Most studies have reported prior achievement as the highest contributing factor in the prediction of student achievement (e.g. Ethington, 1992; Reynolds and Walberg, 1992; Gill and Reynolds, 1999; Fuchs et al., 2000).

Within the context of school effectiveness research (SER), it is generally accepted that a prior achievement variable should be included in the model (for example Willms, 1992; Sammons et al., 1996). Consequently, a vast majority of SER adjusts for at least prior achievement (Gray et al., 1995 & 1996; Hill and Rowe, 1998). However, some debate exists regarding the appropriateness of using a prior achievement variable as a control in SER. This debate mainly arises when prior achievement data are collected proximal to the point at which the school effects are measured (Preece, 1989; Cuttance, 1985) or if the data are collected after a period of study in the same school (Sammons et al., 1996). This debate also arises when it is thought to be difficult to obtain reliable prior achievement information especially for studies conducted at the points of entry to primary or secondary schools (Teddlie et al., 2001). In such cases, those opposed to the use of a prior achievement variable argue that control for this

variable is likely to lead to a reduction in the estimate of the magnitude of school effects and could factor out some variance due to school effects.

Under the above circumstances, it appears that it would be inappropriate to control for prior achievement if the aim of the study were to estimate the absolute magnitudes of the effects of schooling and the absolute variance due to school effects. However, if the focus of the study were to estimate the relative magnitude of school effects and the relative variance due to school effects for a given substantial duration of learning within the same school, then it would appear appropriate to control for prior achievement.

For the current study, the question being asked is, 'By how much has the school contributed to the student's achievement within the two-year period? With this question in mind, it appears appropriate to factor out any school effects and variance due to school effects outside the period of interest. It is considered logical to assume that the duration of time between Grades 3 and 5 (two years) is ample time for schools to have made a substantial impact on student achievement because a considerable amount of teaching should have taken place within the two years.

Socioeconomic status

Studies in Australia and overseas agree that socioeconomic status has a significant influence on student achievement, with students from higher status homes doing better than students from lower status homes (e.g. Porter, 1980; Ainley et al., 1990; Brewer, 1998; Coley, 2002).

Keeves (1995; p.23) reported that the results of the IEA studies with respect to the status of the home indicate that "measures of the socioeconomic status of the home are positively related to students achievement in all countries, at all age levels and for all subjects areas". There is also clear evidence that the strength of association between socioeconomic background and student performance varies from country to country (Lokan et al., 2001; OECD, 2001; Willms and Somers, 2001) and over time (Keeves and Saha 1992).

Ainley et al. (1995) found the correlation between individual socioeconomic background and students' achievement to be lower at the primary school level than at the secondary school level in Australia. In addition, evidence is available in the Australian data from the PISA study to the effect that the relationship between socioeconomic background and numeracy is not as strong as the relationship between socioeconomic background and literacy (Lokan et al., 2001). Moreover, junior secondary data collected in 1975, 1989 and 1995 (Marks and Ainley, 1997) and in 1975, 1995 and 1998 (Rothman, 2003) suggest that the influence of socioeconomic background on numeracy and literacy achievement may be declining in Australia.

In South Australia, a School-card is used to identify students from low economic status, and disadvantaged schools are identified by the proportion of students who possess school-cards. Rothman (1998) reported that in 1997 non-school cardholders were found to have achieved at higher levels than school cardholders in most areas of learning (except English at Grade 3 and below) in government primary schools in South Australia.

For the current study, there are no data on the socioeconomic status (School-card) of the student at the individual-level but this information is available at the school-level. However, the lack of this information at the student-level should not be seen as a major setback in this study. This is because, for both numeracy and literacy, a multivariate analysis of data on Grade 9 students participating in the Longitudinal

Survey of Australian Youth (LSAY) study showed that "socioeconomic status at the individual level has minimal influence on academic achievement, but at the school level it has much greater influence" (Rothman and McMillan, forthcoming; p.25). Moreover, at the Grade 5 primary school level in Australia, VQSP[6] study found that the correlation between achievement level and SES at the student level was small (0.20) but extremely strong (0.90) at the school level (Hill and Rowe, 1996; p. 17). Consequently, Hill and his colleague concluded that SES at the school-level has greater influence on student achievement than at the individual-level in Australia (Hill and Rowe, 1996; p. 17).

It should also be borne in mind that despite the importance of socioeconomic status as a predictor of student achievement, obtaining accurate information on socioeconomic background of the student is usually difficult. The difficulty arises because gathering socioeconomic information basically involves probing matters that most families consider private and therefore young students may not have the information, which means that parents have to be approached for information as well as approval (Bourke, 1998). Because of this problem, it was not possible to obtain information on the socioeconomic status of the individual students included in this study.

Transience, mobility

There are a number of studies that have attempted to examine the influence of transience or mobility on academic achievement. Most studies indicate that mobility has negative effects on student progress in school (Brent and Diobilda, 1993; Rumberger and Larson, 1998; Reynolds and Wolfe, 1999; Temple and Reynolds, 1999; Wright, 1999).

In Australia, a variety of evidence suggests that students change school frequently (Blane, 1985; Kings, 1985; Rahmani, 1985; Mills, 1986; Fields, 1995, 1997a&b). Fields (1995) estimated that about 100,000 Australian children relocated and changed schools every year with many of them relocating several times during their school years. However, there has been relatively little research that examines the educational consequences of student mobility in Australia.

Nevertheless, Fields (1995) found mobile students in Australia experience both academic and social difficulties. In addition, Hill (1996) reported that a major study (School Global Budget Research Project) identified transience as a powerful predictor of school learning in Australia.

In this study and within the context of value added measurement, the progress made by pupils who move between schools would not be due to the efforts of one school alone. Consequently, an appropriate value added score can only be calculated for students who remain in the same school over the study period. Nevertheless, in this study, transience is included in a separate model to investigate its influence on student achievement and its influence on school effects.

Summary

The initial sections of this chapter provide background information about the main concepts of the theories employed to analyze the tests in this study and the equating methods employed to bring the tests onto common scales. The one-parameter IRT (Rasch) model procedures are chosen for calibration, equating and scoring of the tests because the model is considered to be the most robust and easiest to employ.

[6] Victoria Quality School Project

However, before making a decision whether or not to exclude any item from further analysis, a more careful examination is required of the properties of the suspect item using the Rasch model and CTT concepts. The main procedure selected to equate tests in this study is concurrent equating because research studies have shown that this procedure, when compared to alternative procedures, provides a more consistent and stronger measure of the two sets of items and persons being equated (Morrison and Fitzpatrick, 1992; Mohandas 1996).

The later sections of this chapter summarizes what past studies have said or found regarding school effectiveness issues that are of interest in this study. Based upon earlier research, it is evident that Carroll's (1963) model of school learning is a good starting point in developing a school effectiveness model. It is also evident that such a model should include several levels, such as student-level, class-level and school-level.

In modelling for school effectiveness in this study, it should be borne in mind that there are no class-level data, and there are no measures of individual school policy and practices. However, based on the theoretical and statistical model for estimating school effects proposed by Raudenbush and Willms and (1995), which itself is based on Carroll's model of school learning (Carroll, 1963), it can be understood that modelling for school effectiveness is possible in this study.

The next chapter describes the instruments used to collect data, the data sets available for this study and construction of variables from these data sets.

3

Instruments, Data Sets and Data Preparation

A clear understanding of the instruments used to collect the data as well as the structure and the nature of data used in this study are essential if the restrictions of the study are to be understood and appreciated. Consequently, this chapter describes the instruments, the data sets available and the construction of variables from these data sets. The variables described in this chapter are used in the subsequent analyses of the study to identify factors influencing student achievement on the BST.

Instruments used in the study

The tests involved in the BST are developed by the staff of the Basic Skills Testing Unit, Assessment and Reporting Directorate in the New South Wales Department of School Education in consultation and collaboration with the staff of the Curriculum Division Unit of DETE in South Australia. During the process of test development and before the tests are administered they are field tested in schools in another State of Australia. The staff of DETE carry out administration of the tests in South Australia with the assistance of the principals and the class teachers of the participating schools in the State.

The Basic Skills Testing Program (BSTP) instruments always consist of three major sections: (a) student questionnaire, (b) Literacy test, and (c) Numeracy test. A brief description of each of these three instruments of the BSTP is given below.

Student questionnaire

The student questionnaire used in the BSTP at the Grade 3 level is the same one used at the Grade 5 level. At the Grade 3 level, the questionnaire is administered before the Numeracy test while at the Grade 5 level it is administered before the Literacy test.

At both grade levels, the students are required to fill in the questionnaire before proceeding to the other sections of the instrument. The questionnaire contains items that require students to provide information about their gender, age, race, language

spoken in the home, and whether born in Australia or length of stay in Australia. The students are also required to indicate the name as well as the code of their school.

Figure 3.1 presents the items included in the student questionnaire at both grade levels. From Figure 3.1, there is no doubt that the primary aim of the questionnaire is to identify some aspects of the students' home background, together with the students' age and sex.

Print your name here: _____

(First Name) (Last Name)

Print the name of your school here: _____ School Code ☐☐☐☐

Region Code ☐☐

1. Are you a boy or a girl

- ☐ boy
- ☐ girl

2. How old are you? ¥

- ☐ 7 or younger
- ☐ 8
- ☐ 9 or older

3. Are you an Aboriginal person or a Torres Strait Islander person?

- ☐ yes
- ☐ no

4. Does anyone use a language other than English in your home?

- ☐ yes
- ☐ no

5. How often do you speak English in your home?

- ☐ never
- ☐ sometimes
- ☐ usually
- ☐ always

6. How many years have you lived in Australia?

- ☐ 1 or 2
- ☐ 3 or 4
- ☐ 5 or 6
- ☐ more than 6
- ☐ born in Australia

¥For Grade 5 students, the options to this item were:

- ☐ 9 or younger
- ☐ 10
- ☐ 11 or older

Figure 3.1 Student questionnaire, 1995 to 2000

Numeracy tests

A vast majority of the items in the Numeracy tests at both grade levels are multiple choice items, and mostly have four options. However, there are a few open-ended items that are included in the tests especially in the Grade 5 test.

Whatever the type of item, the students are required to indicate their responses in the question paper. For most of the multiple-choice items, blank bubbles are provided next to each option of the item, and the students are asked to colour in the bubble next to the correct answer. For the open-ended items, the students are asked to write down their responses in the blank spaces provided in the question paper. The items in the Numeracy tests always cover three areas of numeracy, namely: Number, Measurement and Space. Table 3.1 presents the items included in each of the three areas in the 1994 to 2000 Numeracy tests for Grade 3 and Grade 5. It is necessary to mention here that the details for the 1994 tests are provided because some portion of the data used in this study was obtained from New South Wales and part of that data contained information that was collected using the 1994 BSTP instruments. The details of the information contained in the data obtained from New South Wales are provided later in this chapter.

The item numbers shown in Table 3.1 are the actual numbers that appeared in the question papers. In the table, the figures in parenthesis are the subtotal of the items in each of the content areas while the figures in bold are the total numbers of items in the test. For example, Table 3.1 shows that the 1994 Numeracy test for Grade 3 had a total of 32 items (7-Space, 14-Number and 11-Measurement), and for Grade 5 the test had total of 44 items (14-Space, 17-Number and 13-Measurement).

A study of the factor structure and the scaling characteristics of the BST carried out by Hungi (1997) using the 1995 BSTP data found that the factor scores for Space, Number and Measurement sub-scales do not differ greatly. Consequently, the study established that it is appropriate to equate and calculate scores for Numeracy but only for curriculum purposes to calculate scores for Space, Number and Measurement. The current scoring practice in the BSTP provides each student with four scores from the Numeracy test, one score on each of the three sub-scales (Space, Number and Measurement) and a total score on a single Numeracy scale.

Literacy tests

The Literacy tests at both grade levels always consist of two sub-tests: (a) the Language sub-test, and (b) the Reading sub-test. At the Grade 3 level, the Language sub-test is the first part of the Literacy test while at the Grade 5 level the Reading sub-test is the first part. Table 3.2 presents the number of items included in the Reading as well as the Language sub-tests in the 1994 to 2000 Literacy tests. For example, Table 3.2 shows that the 1994 Literacy test for Grade 3 had a total of 59 items (33-Reading and 26-Language), while the test for Grade 5 had 81 items (46-Reading and 35-Language), and so on.

Like the items in the Numeracy tests, the majority of the items in the Literacy tests at both grade levels are commonly multiple choice items with four options, and the students are required to indicate their responses by colouring in the bubbles next to the correct answer.

In general, the Language sub-tests usually consist of items that cover at least four areas of the English language, (namely Vocabulary, Spelling, Punctuation, and Structure) and the Reading sub-tests always consist of items that require the students to go through some reading materials that are provided.

Table 3.1 Space, Number and Measurement items included in the Numeracy test in 1994 to 2000

Content	1994		1995		1996		1997		1998		1999		2000	
Area	Items	Subtotal	Items	Subtotal	Items	Subtotal	Items	Subtotal	Items	Subtotal	Items	Subtotal	Items	Subtotal
Grade 3														
Space	2, 8, 10, 17, 21, 25, 27.	*(7)*	12, 15, 22, 25, 31.	*(5)*	9, 10, 14, 20, 24, 29.	*(6)*	1, 2, 7, 18, 20, 29.	*(6)*	1, 9, 15, 16, 21, 26, 27, 29, 32.	*(9)*	2, 10, 16, 21, 22, 27, 29, 30, 32.	*(9)*	3, 8, 20, 26, 27, 30, 31, 34.	*(8)*
Number	4, 5, 7, 9, 11, 12, 14, 16, 18, 20, 22, 26, 30, 31.	*(14)*	1, 4, 6, 7, 8, 9, 10, 11, 13, 14, 17, 19, 23, 26, 27, 28, 29, 30, 32.	*(19)*	1, 3, 6, 7, 8, 11, 12, 13, 15, 17, 18, 19, 22, 25, 27, 28, 30, 31, 32.	*(19)*	6, 8, 9, 11, 13, 15, 16, 17, 19, 21, 23, 24, 25, 26, 27, 30, 31, 32.	*(18)*	2, 4, 5, 6, 7, 8, 10, 11, 13, 17, 18, 22, 24, 25, 28, 31.	*(16)*	1, 4, 5, 7, 8, 9, 12, 13, 14, 17, 19, 20, 23, 24, 31, 34, 35.	*(17)*	1, 5, 6, 7, 9, 12, 14, 15, 16, 17, 19, 21, 23, 24, 25, 33, 35.	*(17)*
Measurement	1, 3, 6, 13, 15, 19, 23, 24, 28, 29, 32.	*(11)*	2, 3, 5, 16, 18, 20, 21, 24.	*(8)*	2, 4, 5, 16, 21, 23, 26.	*(7)*	3, 4, 5, 10, 12, 14, 22, 28.	*(8)*	3, 12, 14, 19, 20, 23, 30.	*(7)*	3, 6, 11, 15, 18, 25, 26, 28, 33.	*(9)*	2, 4, 10, 11, 13, 18, 22, 28, 29, 32.	*(10)*
Total		**32**		**32**		**32**		**32**		**32**		**35**		**35**
Grade 5														
Space	10, 11, 13, 14, 17, 21, 22, 26, 28, 29, 32, 33, 41, 44.	*(14)*	1, 2, 5, 6, 7, 12, 13, 27, 28, 29, 37, 39, 41.	*(13)*	2, 4, 5, 7, 8, 13, 14, 15, 19, 27, 35, 37, 38, 42, 48.	*(15)*	5, 7, 8, 9, 11, 20, 26, 31, 32, 33, 34, 35, 39, 40, 44, 48.	*(16)*	5, 8, 11, 13, 15, 16, 19, 24, 26, 28, 32, 33, 34, 38, 43, 48.	*(16)*	2, 6, 8, 16, 17, 20, 26, 27, 30, 31, 32, 42, 44, 45, 46, 48.	*(16)*	2, 8, 11, 17, 22, 24, 25, 26, 27, 31, 32, 38, 45, 46, 47.	*(15)*
Number	1, 2, 3, 5, 6, 7, 12, 15, 23, 24, 25, 31, 34, 35, 36, 39, 43.	*(17)*	3, 4, 8, 9, 10, 11, 14, 17, 18, 20, 24, 25, 30, 31, 34, 35, 36, 38, 42, 43, 45.	*(21)*	1, 3, 6, 12, 20, 21, 24, 25, 28, 29, 30, 32, 33, 41, 44, 46, 47.	*(17)*	1, 2, 13, 16, 17, 19, 21, 24, 29, 30, 36, 37, 38, 41, 42, 43, 45, 47.	*(18)*	1, 2, 4, 7, 9, 10, 12, 17, 18, 20, 25, 29, 35, 37, 39, 40.	*(16)*	1, 4, 5, 7, 9, 12, 13, 19, 21, 23, 25, 29, 33, 34, 38, 41.	*(16)*	1, 5, 6, 7, 9, 12, 14, 18, 19, 28, 29, 33, 36, 40, 43, 44.	*(16)*
Measurement	4, 8, 9, 16, 18, 19, 20, 27, 30, 37, 38, 40, 42.	*(13)*	15, 16, 19, 21, 22, 23, 26, 32, 33, 40, 44, 46, 47, 48.	*(14)*	9, 10, 11, 16, 17, 18, 22, 23, 26, 31, 34, 36, 39, 40, 43, 45.	*(16)*	3, 4, 6, 10, 12, 14, 15, 18, 22, 23, 25, 27, 28, 46.	*(14)*	3, 6, 14, 21, 22, 23, 27, 30, 31, 36, 41, 42, 44, 45, 46, 47.	*(16)*	3, 10, 11, 14, 15, 18, 22, 24, 28, 35, 36, 37, 39, 40, 43, 47.	*(16)*	3, 4, 10, 13, 15, 16, 20, 21, 23, 30, 34, 35, 37, 39, 41, 42, 48.	*(17)*
Total		**44**		**48**		**48**		**48**		**48**		**48**		**48**

Table 3.2 Numbers of Reading and Language items in the 1994 to 2000 Literacy tests

	Grade 3			Grade 5		
Occasion	Reading	Language	Total	Reading	Language	Total
1994	33	26	**59**	46	35	**81**
1995	32	25	**57**	45	34	**79**
1996	34	25	**59**	46	34	**80**
1997	33	25	**58**	47	36	**83**
1998	34	27	**61**	47	36	**83**
1999	35	28	**63**	46	38	**84**
2000	35	27	**62**	47	36	**83**

For both grade levels, reading materials are provided in the form of a small magazine that is very colourful. These reading materials usually consist of texts, pictures and diagrams that provide information about various things. In order for the students to answer the reading items, they are first instructed to read the material in a specific section of the magazine that corresponds to a set of items in the question paper.

The study of the factor structure and the scaling characteristics of the BST carried out by Hungi (1997) using the 1995 BSTP data established that it is appropriate to equate and calculate scores for Reading and Language as well as a single score for Literacy. The current scoring practice in the BSTP provides each student with the three scores from the Literacy test, that is, one score on each of the two sub-scales (Reading and Language) and a total score on a single Literacy scale.

For a particular testing occasion, a few common items are included in each sub-test of the BST with the purpose of comparing performance between the Grade 3 and Grade 5 levels. Tables 3.3 and 3.4 present lists of the common items included in the 1994 to 2000 Numeracy and Literacy tests respectively. The item numbers shown in the two tables are the actual numbers that appeared in the question papers. For Example, Table 3.3 shows that in the 1994 Numeracy tests, Item 2 in the Grade 3 test was the same as Item 11 in the Grade 5 test, and that Item 8 in the Grade 3 test was the same as Item 10 in the Grade 5 test. Likewise, Table 3.4 shows that in the 1994 Language sub-tests of the Literacy tests, Item 7 in the Grade 3 test was the same as Item 8 in the Grade 5 test, and that Item 33 in the Grade 3 Reading sub-tests was the same as Item 20 in the Grade 5 Reading sub-test. Generally, the common items in the BST are placed in closely similar positions in the two tests.

In Table 3.4, the numbers given in parenthesis are the subtotal of the common items included in each of the sub-tests of the Literacy test. The figures shown in bold in Tables 3.3 and 3.4 are the total numbers of common items in the Grade 3 and Grade 5 Numeracy and Literacy tests for the particular testing occasion. For example, Table 3.3 shows that the 1994 Numeracy tests had nine common items, and Table 3.4 shows that the 1994 Literacy tests had 15 common items (7-Language and 8-Reading).

Data sets

This study uses three main sets of secondary data, namely (a) South Australia BSTP data, (b) school information data obtained from DETE in South Australia, and (c) equating data sets obtained from Department of School Education in New South Wales. The following three sub-sections provide brief descriptions of the information contained in each of these three data sets.

Table 3.3 Common items in the 1994 to 2000 Grades 3 and 5 Numeracy tests

1994		1995		1996		1997		1998		1999		2000	
Grd 3	Grd 5	Grd 3	Grd 5	Grd 3	Grd 5	Grd 3	Grd 5	Grd 3	Grd 5	Grd 3	Grd 5	Grd 3	Grd 5
2	11	5	22	8	3	4	4	2	2	2	2	2	16
8	10	9	36	9	4	7	9	3	3	3	3	4	4
10	33	13	17	10	15	10	12	7	7	4	4	5	5
17	32	15	39	15	24	11	13	11	9	8	7	6	6
18	12	18	33	17	25	12	14	12	14	11	11	7	7
23	37	19	43	21	23	14	15	15	15	12	12	8	8
25	44	20	32	22	32	15	16	16	19	13	13	10	10
27	41	24	40	23	36	17	19	17	17	15	15	12	12
28	27	25	41	24	35	18	20	19	21	16	16	13	13
				25	33	21	17	21	16	18	18	16	18
						22	18			21	20	17	19
												18	20
	9		**9**		**10**		**11**		**10**		**11**		**12**

Table 3.4 Common items in the Grades 3 and 5 Literacy tests

1994		1995		1996		1997		1998		1999		2000	
Grd 3	Grd 5	Grd 3	Grd 5	Grd 3	Grd 5	Grd 3	Grd 5	Grd 3	Grd 5	Grd 3	Grd 5	Grd 3	Grd 5
Language													
7	8	1	1	8	11	12	13	9	11	11	9	13	7
8	9	2	2	9	12	13	14	10	12	12	10	14	8
9	10	3	3	10	13	14	15	11	13	13	11	15	9
10	11	4	4	11	14	15	16	12	14	14	12	16	10
11	12	5	5	12	15	16	17	13	15	15	13	17	11
12	13	6	6	13	16	17	18	14	16	16	14	18	12
13	14	7	7	14	17	18	19	15	17	17	15	19	14
		8	8	15	18			16	18			20	15
		9	9									21	16
		10	10										
		11	11										
	(7)		(11)		(8)		(7)		(8)		(7)		(9)
Reading													
14	14	10	4	6	1	13	9	11	1	5	7	12	7
15	15	11	5	7	2	14	10	12	2	6	8	13	8
16	16	13	6	8	3	15	11	13	3	7	9	14	9
17	21	14	7	9	4	21	21	18	16	8	10	15	10
18	23	18	1	14	9	22	22	19	17	9	11	20	14
30	17	19	2	15	10	23	23	20	18	15	17	21	15
31	18	20	3	16	11	25	24	21	19	16	18	22	16
33	20	25	33	17	12	26	25	23	20	17	19	23	17
		26	34	29	34	27	26	27	26	18	20	24	18
		27	35	30	35	28	27	28	27	20	21	25	19
		28	36	31	36	29	28	29	28	21	22	26	20
				33	37	30	33	30	29	22	23	27	21
						31	34			23	24		
						32	35						
	(8)		(11)		(12)		(14)		(12)		(13)		(12)
Total	**15**		**22**		**20**		**21**		**20**		**20**		**21**

South Australia BSTP data

The information contained in the South Australia BSTP data set has been collected annually as student responses to the BST administered to Grades 3 and 5 students in government schools throughout South Australia since the inception of the BSTP in 1995.

The data available by the year 2000 consist of 106,514 Grades 3 and 5 students in 504 primary schools in South Australia. However, a total of 37,832 (out of the 106,514) students in this data set have data points at both grade levels. This means that the total number of observations is 144,346. Table 3.5 provides a summary of the numbers of students and total observation records in this data set.

In Table 3.5, row R1 records the number of the Grade 3 students included in the data for each of the six occasions and row R2 provides the same information for Grade 5 students. Row R3 of Table 3.5 records the combined number of the Grades 3 and 5 students who took part in the BSTP for each of the six occasions. Row R4 shows the number of Grade 5 students present on each occasion who also have a data point two years earlier, that is, at Grade 3. For example, Table 3.5 shows that 8,018 out of the 10,283 Grade 3 students, who took the tests in 1995, were identified as taking the tests again in 1997 at Grade 5. Row R5 gives the cumulative number of students who have two data points that far. Row R6 gives the cumulative number of students who have two data points in the same school (that is, had not changed school between Grades 3 and 5 grade levels) while row R7 gives the cumulative number for students who have two data points but in two different schools. For example, the table shows that by the year 2000 there were 37,832 students who had two data points and that 32,741 of these students had remained in the same school while 5,091 of these students had changed schools between Grades 3 and 5 levels. Row R8 in Table 3.5 gives the cumulative number of observations recorded that far, for example, the table shows that by 1997 there were 68,136 observations made so far. Finally, Row R9 gives the cumulative number of students who had taken part in the BSTP by that occasion taking into account that some students had two data points as shown in row R4 of Table 3.5.

Table 3.5 Students in the South Australian BSTP data set

		1995	1996	1997	1998	1999	2000
Total Grade 3 Students	R1	10,283	11,095	12,437	12,794	12,550	12,677
Total Grade 5 Students	R2	10,735	11,613	11,973	12,471	12,900	12,818
Total Students	R3	21,018	22,708	24,410	25,265	25,450	25,495
Students matched	R4			8,018	8,972	10,313	10,529
Students matched (Cumulative)	R5			8,018	16,990	27,303	37,832
Students matched same school (Cumulative)	R6			6,898	14,686	23,612	32,741
Students matched changed school (Cumulative)	R7			1,120	2,304	3,691	5,091
Total observations (Cumulative)	R8	21,018	43,726	68,136	93,401	118,851	144,346
Total Students (Cumulative)	R9	21,018	43,726	60,118	76,411	91,548	106,514

Thus, from Table 3.5, it can be noted that in these data some of the students have data points only at one of the grade levels while some have data points at both grade levels. For example, only one data point is available for the Grade 5 students who took the tests in 1995 (N=10,735) and 1996 (N=11,095) because the BSTP had not yet been started in South Australia when those students were in Grade 3. Similarly, only one

data point is available for the Grade 3 students who took the tests in 1999 (N=12,550) and 2000 (N=12,677). In addition, some Grade 3 students have only one data point mostly because they had changed schools between the two grades and moved to schools outside South Australia (or probably to non-government schools in the State). It is also possible that some of these Grade 3 students have only one data point because they could have repeated grades in Grade 3 or Grade 4, or they did not take the test when at Grade 5. Likewise, some Grade 5 students have only one data point mostly because they had joined schools in South Australia in between the two grade levels (probably from schools outside the State or from private schools in the State). Possibly, some of these Grade 5 students have only one data point because they did not take the tests when in Grade 3 or might have repeated either Grade 3 or Grade 4.

There are also some chances that a few students have a single data point because their Grade 3 and Grade 5 data could not be matched due to change of names. It is also likely that a few of the students have a single data point because they had been exempted from the BSTP when they were in Grade 3 or when they were in Grade 5. However, there is no information available to differentiate between the students who have missing data due to mobility or repetition of grades or any other reasons and those who have missing data due to exemption from the test.

Table 3.6 displays a summary of the number of schools included in this data set on each occasion. Row R1 of Table 3.6 shows the number of schools that participated in the BSTP on each of the six occasions. Row R2 shows the number of schools that participated in the BSTP for the first time on that occasion. For example, Table 3.6 shows that there were four schools that participated for the first time in 1997. Row R3 shows the total number of schools that had participated so far by that occasion since the inception of the program in South Australia.

Table 3.6 Schools' participation by occasion

		Occasion					
		1995	**1996**	**1997**	**1998**	**1999**	**2000**
Total participation	R1	489	485	482	474	473	468
First participation	R2	489	6	4	1	1	3
Cumulative participation	R3	489	495	499	500	501	504

Note: A few primary schools in South Australia have closed down each year due to lack of students, thus the general drop in the total number of schools participating in the BSTP in successive years.

Table 3.7 presents a summary of the number of times that data were collected from the schools included in this data set. It should be noted that not all the schools have data for all the six occasions. For example, of the total 504 schools in the data set, only 455 schools have data records for all the six occasions as recorded in the first row of Table 3.7 under column C1. In addition, it should be noted that on some occasions some schools have data only for one of the grade levels and not the other. Consequently, only 426 schools out of the 504 schools have data for both grade levels for all the six occasions as shown in the first row of Table 3.7 under column C2. Similarly, the last row of Table 3.7 under column C2 shows that there are six schools out of the 504 schools that have data collected from only one of the grade levels during the period of the study.

It is evident from the information displayed in Table 3.7 that only 426 out of 504 schools (around 85 per cent) have participated fully in the BSTP since its inception.

Table 3.7 Number of participation times by schools

Number of times	Participation (C1)	Participation in both grade levels (C2)
Six	455	426
Five	10	32
Four	6	6
Three	13	12
Two	8	11
One	12	11
Zero	0	6
Total Schools	**504**	**504**

School information data

The main South Australian BSTP data described above contains no information regarding the schools in the data. For example, the main BSTP data does not contain information about the student's school attendance record, the identity of school-card holders, locality of the school and so on. Such school information is essential in the development of an "explicit theory of good standing" (McPherson, 1993: p.2) for the assessment of the school effects on student achievement in South Australia primary schools. Consequently, this information was obtained from DETE in South Australia. However, it should be noted that the school data available from DETE are mainly for administrative purposes and therefore lack some details. For example, the data give information on the total numbers of school-card holders in each school but do not give any information to facilitate the identification of the school-card holders in the school. Likewise, the data provide just the absenteeism rates per school per academic year but no information to show how many days the individual students were absent from school in the academic year. Consequently, at the student-level, it is not possible to match the information obtained from DETE with the information obtained from the BSTP data.

The following ten items of school information were obtained from DETE for each school that participated in the BSTP from 1995 to 2000:

(a) name of the school,

(b) code number of the school,

(c) the district the school is located,

(d) number of students in the school,

(e) locality of the school — (rural or urban),

(f) distance of the school from Adelaide,

(g) number of school-card holders in the school,

(h) affiliation of the school to the Country Area Program,

(i) average absenteeism rate in the school, and

(j) mobility rate in the school.

The code number of the school used in the BSTP data is the same one used in the school information data obtained from DETE. Therefore, at the school level, there were no problems in matching the two data sets.

New South Wales equating data

In the BSTP there are no items included on more than one testing occasion. Hence, comparison of performance between testing occasions cannot be directly achieved in this study by the use of common item equating.

However, in New South Wales (NSW), all the Basic Skills Tests administered in that State are linked back to the 1996 test either directly (or indirectly through another occasion test) so as to equate tests across occasions. Since the Basic Skills Tests that are used in NSW are the same ones used in South Australia, some equating data were obtained from NSW Department of School Education to help link the tests from the different testing occasions in South Australia.

Thus, NSW equating data consists of groups of Grades 3 and 5 students from NSW who had taken the 1996 tests (or another test directly linked to the 1996 tests) as a trial test a week prior to taking the real test for that occasion. However, the NSW equating contains no information regarding the background of the students and the schools in the data because the information collected using the student questionnaire is not included in the data set. Therefore, it is not possible to tell how many schools are involved in the study, neither is it possible to identify the schools.

Table 3.8 provides a summary of the number of equating students (Groups 1 to 6) in the NSW equating data set on each occasion. For example, Table 3.8 shows that in 1995 (that is, Group 1) there were 1,646 (829 Grade 3 and 817 Grade 5) students from NSW who took the 1994 tests as a trial a week prior taking the real test, that is, the 1995 tests. From Table 3.8, it can be noted that 976 Grade 5 students in NSW took the 1994 tests, as trial tests in 1996, however, it can also be noted that there are no data for the Grade 3 students who took trial tests on the 1996-testing occasion. Therefore, equating of the 1994 Grade 3 tests to the 1996 Grade 3 tests must necessarily rely on the common items included in the 1996 Grades 3 and 5 tests.

Table 3.8 Numbers of Grade 3 and Grade 5 students in the NSW equating data

	Real Test	Trial Test	Grade 3	Grade 5	Total
Group 1	**1995**	1994	829	817	1,646
Group 2	**1996**	1994	Nil	976	976
Group 3	**1997**	1996	939	1,030	1,969
Group 4	**1998**	1997	1,154	1,204	2,358
Group 5	**1999**	1996	1,000	974	1,974
Group 6	**2000**	1999	1,088	1,126	2,214
		Total	**5,010**	**6,127**	**11,137**

Table 3.9 presents the total number of students included in the NSW data set who had taken the tests for a particular occasion either as the real test or the trial test by the year 2000. For example, Table 3.9 shows that by the year 2000, there was a total of 2,093 Grade 3 equating students from NSW who had taken the 1997 test that far. This total (2,093) is obtained by adding up the Grade 3 equating students who had that far taken the 1997 tests, that is, adding up the 939 students from the 1997 testing occasion and the 1154 students from the 1998 testing occasion as given in Table 3.8.

Preparation of the data for analyses

This section describes preparations and coding made to the data before starting the analyses. In particular, the section provides a brief description of the predictor

variables constructed from the South Australia BSTP data and the School Information data sets described above. The procedures followed in the construction of achievement related variables (that is, student scores at Grades 3 and 5 in the Basic Skills Tests) are provided in Chapter 6.

The next three sub-sections describe the student-level variables, the school-level variables and occasion-related variables that are constructed for use in subsequent analyses in this study.

Table 3.9 Participation sizes of the NSW equating students in the BSTP by the year 2000

	Testing Occasion						
	1994	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**
Grade 3	829	829	1,939	2,093	1,154	2,088	1,088
Grade 5	1,793	817	2,980	2,234	1,204	2,100	1,126
Totals	**2,622**	**1,646**	**4,919**	**4,327**	**2,358**	**4,188**	**2,214**

Construction of the student-level variables

The following student-level variables are constructed from the South Australia BSTP data set described above: YEARLEVL, SEX, AGE, ATSI, NESB, HOME, INOZ and TRANS. Information on these variables was obtained from the student questionnaire included in the test booklet. A brief description of these variables is provided below.

YEARLEVL denotes the grade level of the student. In order to reflect the two years difference in study between the two grade levels, a zero (0) is used to indicate a Grade 3 student and a two (2) to indicate a Grade 5 student. The breakdown of the numbers of students who took part in the BSTP by their grade levels has already been provided above (Table 3.1).

Over the six occasions, the data contained a total of 71,836 observations at Grade 3 and 72,510 observations at Grade 5. There were no missing data on this variable because the grade level was pre-printed at the top of the test booklet.

SEX denotes the gender of the student with a zero (0) indicating male and a one (1) indicating female. A breakdown of the numbers (and percentages) of students who took part in the BSTP on each occasion by their gender is presented in Table 3.10. This information was obtained from the students' responses to Item 1 of the questionnaire shown in Figure 3.1.

From Table 3.10, it can be observed that in total, 73,605 (51.0 per cent) observations were obtained from boys and 70,741 (49.1 per cent) observations were obtained from girls over the six occasions. It should be mentioned that over the six occasions a total of 72 students did not indicate their gender and therefore the gender of these students was coded as missing in the original data. However, for the purposes of this study and since only a few students did not indicate their gender, the gender of these students was obtained from looking up their names on the master file and making a judgment with respect to the sex of the student.

AGE denotes the age of the student. This information was obtained from the students' responses to Item 2 of the questionnaire shown in Figure 3.1.

In coding the data, a zero (0) is used to indicate a Grade 3 student who was seven years of age or younger and a Grade 5 student who was nine years of age or younger. Likewise, a one (1) is used to indicate a Grade 3 student who was eight years and a

Grade 5 student who was ten years. A two (2) is used to indicate Grade 3 student who was nine years or older and Grade 5 student who was 11 years or older.

Table 3.10 Students included in the study by their gender, age, race, and English speaking background

	1995		1996		1997		1998		1999		2000		Total	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Grade 3 Gender														
Boys	5,231	50.9	5,623	50.7	6,393	51.4	6,442	50.4	6,534	52.1	6,466	51.0	36,689	51.1
Girls	5,052	49.1	5,472	49.3	6,044	48.6	6,352	49.7	6,016	47.9	6,211	49.0	35,147	48.9
Grade 5 Gender														
Boys	5,452	50.8	5,964	51.4	6,047	50.5	6,446	51.7	6,586	51.1	6,421	50.1	36,916	50.9
Girls	5,283	49.2	5,649	48.6	5,926	49.5	6,025	48.3	6,314	49.0	6,397	49.9	35,594	49.1
Total Gender														
Boys	10,683	50.8	11,587	51.0	12,440	51.0	12,888	51.0	13,120	51.6	12,887	50.6	73,605	51.0
Girls	10,335	49.2	11,121	49.0	11,970	49.0	12,377	49.0	12,330	48.5	12,608	49.5	70,741	49.0
Grade 3 AGE														
7 or younger	127	1.2	138	1.2	135	1.1	130	1	172	1.4	199	1.6	901	1.3
8 years	8,739	85.0	9,388	84.6	10,111	81.3	10,464	81.8	10,509	83.7	10,639	83.9	59,850	83.3
9 or older	1,409	13.7	1,561	14.1	2,155	17.3	2,188	17.1	1,863	14.8	1,836	14.5	11,012	15.3
Missing	8	0.1	8	0.1	36	0.3	12	0.1	6	0.1	3	0.0	73	0.1
Grade 5 AGE														
9 or younger	160	1.5	154	1.3	116	1.0	137	1.1	113	0.9	147	1.2	827	1.1
10 years	9,036	84.2	9,762	84.1	10,171	85.0	10,610	85.1	10,598	82.2	10,556	82.4	60,733	83.8
11 or older	1,539	14.3	1,688	14.5	1,675	14.0	1,715	13.8	2,156	16.7	2,100	16.4	10,873	15.0
Missing	0	0.0	9	0.1	11	0.1	9	0.1	33	0.3	15	0.1	77	0.1
Students of Aboriginal and Torres Islander Origin														
Grade 3 ATSI														
ATSI	379	3.7	421	3.8	519	4.2	564	4.4	472	3.8	499	3.9	2,854	4.0
Non-ATSI	9,874	96.0	10,674	96.2	11,918	95.8	12,205	95.4	12,009	95.7	12,079	95.3	68,759	95.7
Missing	30	0.3	0	0.0	0	0.0	25	0.2	69	0.6	99	0.8	223	0.3
Grade 5 ATSI														
ATSI	284	2.7	326	2.8	429	3.6	485	3.9	525	4.1	562	4.4	2,611	3.6
Non-ATSI	10,378	96.7	11,287	97.2	11,544	96.4	11,962	95.9	12,362	95.8	12,246	95.5	69,779	96.2
Missing	73	0.7	0	0.0	0	0.0	24	0.2	13	0.1	10	0.1	120	0.2
Total ATSI														
ATSI	663	3.2	747	3.3	948	3.9	1,049	4.2	997	3.9	1,061	4.2	5,465	3.8
Non-ATSI	20,252	96.4	21,961	96.7	23,462	96.1	24,167	95.7	24,371	95.8	24,325	95.4	138,538	96.0
Missing	103	0.5	0	0.0	0	0.0	49	0.2	82	0.3	109	0.4	343	0.2
Students of non-English speaking background														
Grade 3 NESB														
NESB	1,354	13.2	1,523	13.7	1,916	15.4	1,770	13.8	1,889	15.1	1,767	13.9	10,219	14.2
ESB	8,914	86.7	9,562	86.2	10,429	83.9	10,987	85.9	10,649	84.9	10,905	86.0	61,446	85.5
Missing	15	0.2	10	0.1	92	0.7	37	0.3	12	0.1	5	0.0	171	0.2
Grade 5 NESB														
NESB	1,456	13.6	1,483	12.8	1,578	13.2	1,812	14.5	1,848	14.3	1,687	13.2	9,864	13.6
ESB	9,277	86.4	10,113	87.1	10,376	86.7	10,642	85.3	10,964	85.0	11,095	86.6	62,467	86.2
Missing	2	0.0	17	0.2	19	0.2	17	0.1	88	0.7	36	0.3	179	0.3
Total NESB														
NESB	2,810	13.4	3,006	13.2	3,494	14.3	3,582	14.2	3,737	14.7	3,454	13.6	20,083	13.9
ESB	18,191	86.6	19,675	86.6	20,805	85.2	21,629	85.6	21,613	84.9	22,000	86.3	123,913	85.8
Missing	17	0.1	27	0.1	111	0.5	54	0.2	100	0.4	41	0.2	350	0.2

A breakdown of the numbers (and percentages) of students who took part in the BSTP on each occasion by their age categories is also presented in Table 3.10. It should be noted from Table 3.10 that a few students (about 0.1 per cent) at both grade levels did not indicate their age category and therefore the age categories of these students are coded as missing.

ATSI denotes the racial background of the student. This information was obtained from the students' responses to Item 3 of the questionnaire shown in Figure 3.1. In data coding, a zero (0) is used to indicate Aboriginal or Torres Strait Islander (ATSI) student and a one (1) is used to indicate non-ATSI student. Table 3.10 also presents a breakdown of the numbers (and percentages) of student included in this study by their racial groups. It should be noted that 343 (about 0.3 per cent) students did not provide information about their racial background and are therefore coded as having missing data.

NESB and **HOME** denote information on the students' English speaking background. Information on Non-English Speaking Background (NESB) was obtained from the students' responses to Item 4 of the questionnaire shown in Figure 3.1, and information on speaking English at home (HOME) was obtained from responses to Item 5 of the questionnaire. Hence, the two variables NESB and HOME are alternative versions of the same measure and therefore, should not be included simultaneously in any model tested in the analysis of data to avoid problems associated with suppressor variables (Keeves, 1997).

In coding the data, a zero (0) is used to indicate a student from a non-English speaking background (following a 'yes' response to Item 4 of the student questionnaire) and a one (1) to indicate a student from an English speaking background (following a 'no' response to Item 4 of the student questionnaire). For HOME (Item 5 of the student questionnaire), a zero (0) is used to indicate a 'never' response, a one (1) to indicate 'sometimes', a two (2) to indicate 'usually', and a three (3) to indicate 'always'.

A breakdown of the numbers (and percentages) of students who took part in the BSTP on each occasion by their English speaking background (NESB) is presented in Table 3.10, and a breakdown by their speaking English at home (HOME) is given in Table 3.11. It should be noted from Tables 3.10 and 3.11 that 350 students did not respond to the NESB item and 413 students did not respond to the HOME item.

INOZ denotes a student's length of living in Australia, that is, migrant status. The INOZ information was obtained from the students' responses to Item 6 of the student questionnaire shown in Figure 3.1 above. The responses to Item 6 are coded so as to reflect the duration of living in Australia from low (1 or 2 years) to high (born in Australia). Consequently, data codes for the item ranges from a zero (indicating a student who has lived in Australia for one to two years) to a four (indicating a student born in Australia).

A breakdown of the numbers (and percentages) of students who took part in the BSTP on each occasion by INOZ is presented in Table 3.11. From the Table 3.11, it should be noted that a total of 367 students did not indicate their length of stay in Australia.

TRANS denotes a student's transience, that is, whether or not the student had changed school between Grades 3 and 5 levels. In this study, the availability of the TRANS information was expected for a maximum of 46,609 students only. That is, the TRANS information was expected for the students who took the tests at Grade 3 in 1995 to 1998 (N=46,609), and not for the students who took the tests at Grade 3 in 1999 and 2000 (N=25,227).

Table 3.11 Students included in the study by their English speaking background, and length of stay in Australia

	1995		1996		1997		1998		1999		2000		Total	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%
English Spoken in the home														
Grade 3 HOME														
Never	323	3.1	248	2.2	283	2.3	325	2.5	270	2.2	271	2.1	1,720	2.4
Sometimes	583	5.7	628	5.7	687	5.5	704	5.5	676	5.4	666	5.3	3,944	5.5
Usually	645	6.3	745	6.7	958	7.7	841	6.6	905	7.2	865	6.8	4,959	6.9
Always	8,601	83.6	9,474	85.4	10,509	84.5	10,861	84.9	10,651	84.9	10,819	85.3	60,915	84.8
Missing	131	1.3	0	0.0	0	0.0	63	0.5	48	0.4	56	0.4	298	0.4
Grade 5 HOME														
Never	197	1.8	142	1.2	162	1.4	153	1.2	134	1.0	159	1.2	947	1.3
Sometimes	468	4.4	444	3.8	480	4.0	512	4.1	562	4.4	527	4.1	2,993	4.1
Usually	726	6.8	694	6.0	723	6.0	784	6.3	840	6.5	842	6.6	4,609	6.4
Always	9,291	86.6	10,333	89.0	10,608	88.6	10,997	88.2	11,345	88.0	11,272	87.9	63,846	88.1
Missing	53	0.5	0	0.0	0	0.0	25	0.2	19	0.2	18	0.1	115	0.2
Total HOME														
Never	520	2.5	390	1.7	445	1.8	478	1.9	404	1.6	430	1.7	2,667	1.9
Sometimes	1,051	5.0	1,072	4.7	1,167	4.8	1,216	4.8	1,238	4.9	1,193	4.7	6,937	4.8
Usually	1,371	6.5	1,439	6.3	1,681	6.9	1,625	6.4	1,745	6.9	1,707	6.7	9,568	6.6
Always	17,892	85.1	19,807	87.2	21,117	86.5	21,858	86.5	21,996	86.4	22,091	86.7	124,761	86.4
Missing	184	0.9	0	0.0	0	0.0	88	0.4	67	0.3	74	0.3	413	0.3
Length of period living in Australia														
Grade 3 INOZ														
1 or 2 Years	109	1.1	127	1.1	112	0.9	128	1.0	150	1.2	151	1.2	777	1.1
2 or 4 Years	159	1.6	140	1.3	145	1.2	170	1.3	166	1.3	159	1.3	939	1.3
5 or 6 Years	155	1.5	144	1.3	144	1.2	123	1.0	178	1.4	164	1.3	908	1.3
More than 6 years	159	1.6	182	1.6	159	1.3	139	1.1	293	2.3	264	2.1	1,196	1.7
Born in Australia	9,660	93.9	10,502	94.7	11,877	95.5	12,208	95.4	11,674	93.0	11,831	93.3	67,752	94.3
Missing	41	0.4	0	0.0	0	0.0	26	0.2	89	0.7	108	0.9	264	0.4
Grade 5 INOZ														
1 or 2 Years	83	0.8	105	0.9	120	1.0	144	1.2	118	0.9	142	1.1	712	1.0
3 or 4 Years	137	1.3	114	1.0	107	0.9	112	0.9	121	0.9	105	0.8	696	1.0
5 or 6 Years	183	1.7	190	1.6	156	1.3	142	1.1	135	1.1	149	1.2	955	1.3
More than 6 years	259	2.4	340	2.9	331	2.8	286	2.3	290	2.3	252	2.0	1,758	2.4
Born in Australia	10,004	93.2	10,864	93.6	11,259	94.0	11,775	94.4	12,223	94.8	12,161	94.9	68,286	94.2
Missing	69	0.6	0	0.0	0	0.0	12	0.1	13	0.1	9	0.1	103	0.1
Total INOZ														
1 or 2 Years	192	0.9	232	1.0	232	1.0	272	1.1	268	1.1	293	1.2	1,489	1.0
2 or 4 Years	296	1.4	254	1.1	252	1.0	282	1.1	287	1.1	264	1.0	1,635	1.1
5 or 6 Years	338	1.6	334	1.5	300	1.2	265	1.1	313	1.2	313	1.2	1,863	1.3
More than 6 years	418	2.0	522	2.3	490	2.0	425	1.7	583	2.3	516	2.0	2,954	2.1
Born in Australia	19,664	93.6	21,366	94.1	23,136	94.8	23,983	94.9	23,897	93.9	23,992	94.1	136,038	94.2
Missing	110	0.5	0	0.0	0	0.0	38	0.2	102	0.4	117	0.5	367	0.3

In the BSTP, students are given new identification numbers on each testing occasion. Therefore, two different identification numbers are given to the same student who has two data points, one identification number at Grade 3 and the other at Grade 5. However, the same code numbers are used to identify the schools participating in the BSTP on the different testing occasions. Therefore, using the school codes and the students' names, it was possible to match manually the students who were expected to have two data points especially those who remained in the same school over the two-year period. However, matching of the students who had changed school in between the two grade levels relied mainly on the names of the students together with the

correct combination of the information given by the students in the questionnaire about their background at Grade 3 and at Grade 5.

About 81.2 per cent (N=37,832) of the 46,609 students were successfully matched, which left 8,777 students (about 18.8 per cent) unmatched. Some 32,741 of the matched students had remained in the same schools while 5,091 students had changed schools in between the two grade levels. A summary of the numbers of students matched on each testing occasion has already been given above in Table 3.5.

In general, the percentage (81.2) of students matched in this study is considered substantial. This is because it is considered reasonable to assume that the 8,777 students (or 18.8 per cent) could not be matched mainly because they had changed schools between the two grades and moved to schools outside South Australia (or possibly to non-government schools in the State). In addition, the number of students who remained unmatched is considered to be low especially if it is borne in mind that it is a total for six testing occasions, 1995 to 2000. Furthermore, several studies have suggested that Australian students change school frequently (Blane, 1985; Mills, 1986; Fields, 1995 and 1997a&b). Moreover, Fields (1995) estimated that about 100,000 Australian children relocate and change schools every year with many of them relocating several times during their school year, and the South Australian proportion (9.0 per cent) could give 9,000 transience students per year across all eight primary school grades or approximately 1,000 for each grade. However, it should not be forgotten that it is also possible that some of students might have failed to be matched because of other reasons such as repetition of grades in Grade 3 or Grade 4, or failure to take the test when at Grade 5.

In coding the data, a zero (0) is used to indicate a student who had remained in the same school over the two-year duration, a one (1) to indicate a student who had changed schools, and a nine (9) to indicate a student whose transience information is unknown.

Construction of the school-level variables

The school-level variables constructed from the South Australia BSTP and School Information data sets described above can be grouped into two categories: (a) student related school-level variables (b) student free school-level variables.

Student-related school-level variables

Variables under this category are formed (a) by aggregating the student-level variables described above, and from (b) school information data set obtained from DETE in South Australia.

Aggregated variables

Aggregating the student-level variables forms these variables at the school-level. A number of studies have shown that school-level aggregate values of an individual level variable could be significantly related to achievement even after controlling for the variable at the individual level (Harker and Nash, 1996).

Aggregating the student-level variables described above within each testing occasion forms a total of seven variables. The aggregated variables are namely SEX_1, AGE_1, ATSI_1, NESB_1, HOME_1, INOZ_1 and TRANS_1. These variables do not indicate the overall composition of students in the schools included in this study but the composition of the students who took the tests in those schools for a particular testing occasion. Importantly, these variables represent the changing school context and should not be confused with average school context variables (AGE_2, ATSI_2,

NESB_2, HOME_2, INOZ_2 and TRANS_2), that are formed by aggregating the student-level variables at the school-level across all the testing occasions. In this study, codes of changing school context variables have _1 suffix at the end while codes of the average school context variables have _2 suffix at the end.

When interpreting the aggregated variables, it should be remembered how each student-level variable was originally coded. Hence, the variable SEX_1 denotes the proportion of girls (since boy=0, girl=1), AGE_1 denotes the average age of the students, ATSI_1 denotes the proportion of non-ATSI students, and NESB_1 denotes the proportion of English speaking students in a school on the particular occasion. Likewise, the variable HOME_1 denotes the average speaking English at home, INOZ_1 denotes the average duration of living in Australia, and TRANS_1 denotes the proportion of Grade 5 who are newcomers in a school. Aggregating the above variables across all the testing occasions gives SEX_2, AGE_2, ATSI_2, NESB_2, HOME_2, INOZ_2 and TRANS_2 respectively.

An eighth student-related school-level variable, YR35PPT, is formed by totalling the number of students in the school who took the test on each occasion. Hence, the variable YR35PPT denotes participation size of the school. It was not possible to work out the participation ratio of individual schools because DETE did not provide the numbers of students who did not participate in the program for the individual schools.

Variables from school information data set

There are three student-related school-level variables formed from the school information data set obtained from the Department for Education, Training and Employment (DETE) in South Australia. These variables are PSCARD, MOBILITY, and ABSENT. These variables are aggregated across all the testing occasions to obtain PSCARD_2, MOBILI_2 and ABSENT_2. A brief description of each of these three variables is given below.

PSCARD denotes the proportion of school-card holders in the whole school within a particular testing occasion while PSCARD_2 denotes the same but aggregated across all testing occasions. The government provides a School-card to students of low social economic status so that they may obtain concessions in a number of services. Hence, the school-card is an indicator of socioeconomic status (SES) of the family. Consequently, schools with high PSCARD (or PSCARD_2) score have a large proportion of their students from lower SES homes and those schools with a low PSCARD (or PSCARD_2) score have a large proportion of their students from higher SES homes.

MOBILITY denotes the proportion of the students changing school within an academic year for a particular testing occasion while MOBILI_2 denotes this mobility information aggregated across all the testing occasions. Therefore, a higher score for these variables indicates that the school has a higher proportion of mobile students and a lower score indicates that the school has a higher proportion of students who rarely change school.

ABSENT denotes the total number of days students miss schools as a proportion of the total number of days in the academic year for a particular testing occasion while ABSENT_2 denotes this school absenteeism information aggregated across all the testing occasions. Hence, a higher ABSENT (or ABSENT_2) score indicates that the students miss school more often in that school and a lower score indicates a higher school attendance rate by the students in that school.

Student-free school-level variables

All the variables in this category are formed from the school information data set obtained from DETE in South Australia. These variables are: SSIZE (also SSIZE_2), METRO, GPODIST, and CAP. The definitions together with brief descriptions of these five variables are given below.

SSIZE denotes the number of students in the school within a particular testing occasion while SSIZE_2 denotes the average number of students in the school over the study period. The difference between the number of students in the largest and the smallest school included in the study is 1,543 students. This range can be considered huge especially when it comes to the interpretation of the regression coefficient of the variable obtained from HLM analyses. Furthermore, descriptive statistics (plus a histogram plot) of SSIZE indicate that the variable is markedly positively skewed (skewness=1.83) and has a relatively large standard deviation (180.15) compared to its mean (238.65). Therefore, due to the nature of the distribution of the variable, a decision was made to transform the variable.

In situations where the raw variable is positively skewed, as is the case with SSIZE, Tabachnick and Fidell (1989) have suggested trying the logarithm, the square root or the inverse data transformation methods. Regarding the selection of the data transformation method to be applied, Tabachnick and Fidell have recommended as follows:

> If you decide to transform, it is important to check that the variable is normally or near-normally distributed after the transformation. Often you need to try first one transformation and then another until you find the transformation that produces skewness and kurtosis values nearest zero, or the fewest outliers. (Tabachnick and Fidell, 1989; p.84)

All the three transformation methods suggested by Tabachnick and Fidell were tried and then the resulting distributions were examined to select the data transformation method that would be appropriate for the variable SSIZE. Figure 3.2 shows the distribution of the variable before (SSIZE) and after the logarithm (SSIZELOG[7]), square root (SSIZESRT)[8] and inverse (SSIZEINV)[9] transformations.

From Figure 3.2, it can be observed that the logarithm transformation and the square root transformation give near normal distributions. However, it can also be observed that more outliers are obtained after the square root transformation than after the logarithm transformation. Therefore, for the purposes of HLM analyses, the logarithm transformation is selected as the method to transform most appropriately the variable SSIZE.

Since large schools after the logarithm transformation still retain high values compared to small schools, it is assumed that the logarithm transformation does not alter the original meaning as far as the direction and the significance of the effects are concerned. Running the same HLM analysis on the raw and transformed data to see whether the results changed can check this assumption simply. However, the magnitudes of the effects would change and it is important to remember the nature of the transformation made to the variable when interpreting the results (Tabachnick and Fidell, 1989).

---

[7] SSIZELOG = $\log_{10}$[SSIZE]
[8] SSIZERT = √SSIZE
[9] SSIZEINV = 1/SSIZE

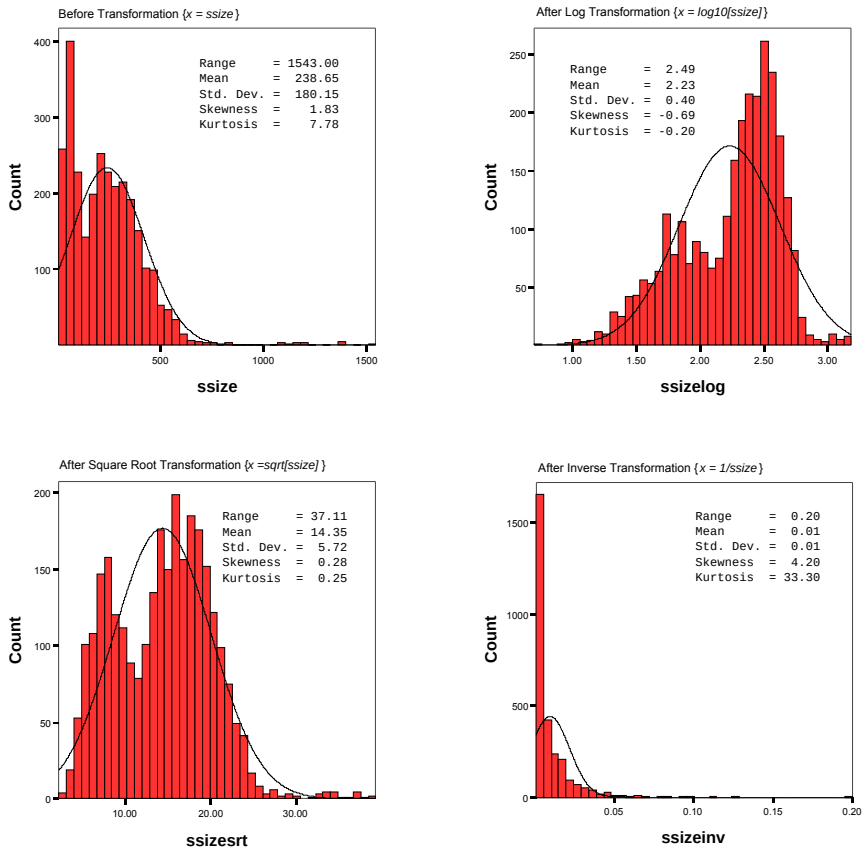


Figure 3.2 Distribution of SSIZE before and after the transformations

METRO denotes the locality of the school, that is, whether rural or urban. In data coding, a zero (0) is used to indicate a country (rural) area school and a one (1) to indicate a metropolitan (urban) area school.

GPODIST denotes the distance of the school from the Adelaide General Post Office (GPO). It is hypothesized that the GPODIST influences school performance because in South Australia it is an indicator of how remote the school is. Adelaide is the only major metropolitan area in South Australia and, therefore, schools that are far from Adelaide are considered remote.

It is worth noting that GPODIST is considered to be an alternative version of METRO. This is because all the urban schools included in this study are within 47 kilometers from Adelaide GPO, and a vast majority (92 per cent) of the rural schools are beyond 47 kilometers from the GPO. Hence, the two variables should not be included simultaneously in any model tested in the analysis of data to avoid problems associated with multicollinearity and suppressor variable relationships (Keeves, 1997).

The descriptive statistics of GPODIST indicated that this variable also required data transformation because it has a range of 1999 kilometers with a mean of 170.74, a standard deviation of 265.68 and a positive skewness value of 3.12. Consequently, the three data transformation methods (that is, logarithm, square root and inverse) that

were suggested by Tabachnick and Fidell (1989) for positively skewed data were also tested for this variable. Figure 3.3 shows the distribution of the variable before transformation (GPODIST) and after the logarithm transformation (GPOLOG[10]), square root transformation (GPOSQRT[11]), and after the inverse transformation (GPOINV[12]).

From Figure 3.3, it can be observed that the logarithm transformation gives the nearest to normal distribution and has fewer outliers compared to the other transformation methods tried. Consequently, the logarithm transformation is selected as the method to transform appropriately the variable GPODIDT for use in the HLM analyses.

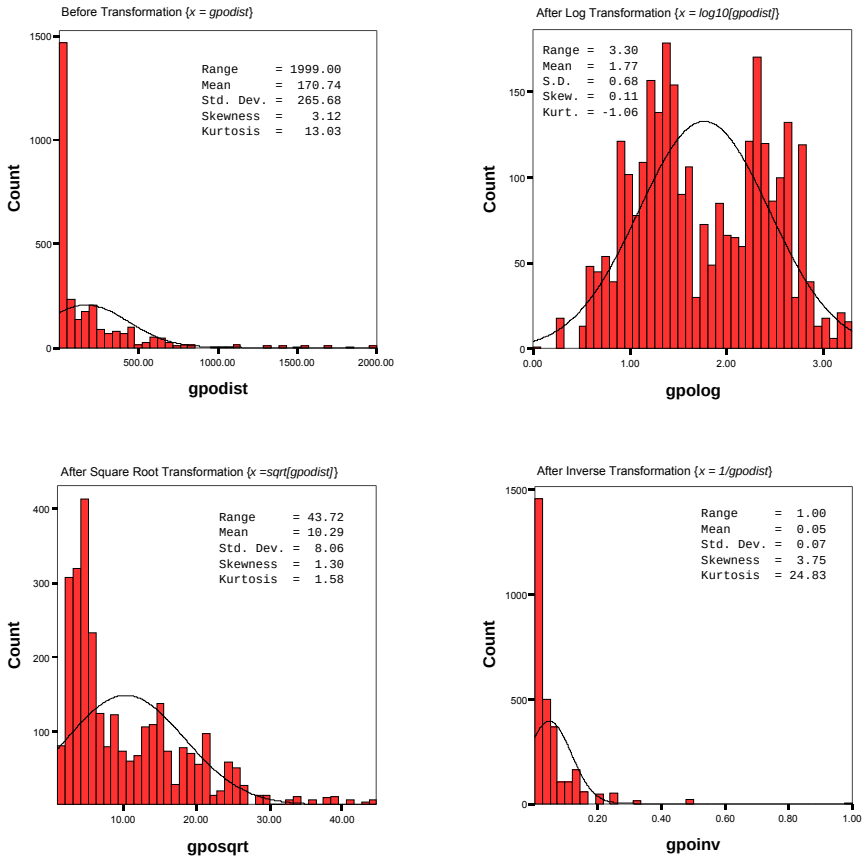


Figure 3.3 Distribution of GPODIST before and after the transformations

CAP denotes affiliation of the school to the Country Area Program for a particular testing occasion. The Federal Government provides financial assistance to schools in the Country Area Program (CAP). The schools are included in the CAP if they are considered financially disadvantaged by the Federal Government and, therefore, CAP is an indicator of the economic status of the school. In coding the data, a one (1) is used to indicate a school that receives CAP funding and, a zero (0) to indicate a school not receiving the funding.

[10] GPOLOG = $\log_{10}$[GPODIST]
[11] GPOSQRT = √GPODIST
[12] GPOINV = 1/GPODIDT

Construction of the occasion-related variables

Six occasion-related dummy variables (OCC1, OCC2, ... , OCC6) are constructed from the South Australia BSTP data set. Each dummy variable indicates the occasion of measurement with OCC1 denoting 1995, OCC2 denoting 1996 and so on. In coding the data, a one (1) is used to indicate data collected on that occasion and a zero (0) is used to indicate data collected on the other five occasions.

Two occasion-related trend variables (OCC and OCCSQD) are also constructed from this data set. OCC is a linear trend variable (coded 1995=0, 1996=1, . . . , 2000=5) and it was, therefore, constructed to facilitate the testing for a linear relationship of the variables between the testing occasions. OCCSQD was constructed by squaring OCC. Hence, OCCSQD is a quadratic trend variable (coded 1995=1, 1996=4, . . . , 2000=36) and it was constructed to permit the testing for a curvilinear relationship of the variables between the six occasions.

It should be noted that in models where the point of reference is the four cohorts of students rather than the six testing occasions, the linear trend variable OCC is coded '1995/1997 cohort' =0, ... '1998/2000 cohort' = 3, and accordingly, OCCSQD codes are 1, 4, 9 and 16.

Summary

Most of the data for this study were obtained from the Department for Education, Training and Employment in South Australia. These data have been collected annually as student responses to the Basic Skills Tests administered to Grades 3 and 5 students in government schools throughout South Australia since the inception of the BSTP in 1995.

The Basic Skills Tests are developed by the staff of the Basic Skills Testing Unit, Assessment and Reporting Directorate in the NSW Department of School Education in consultation and collaboration with the staff of the Curriculum Division Unit of DETE in South Australia. In South Australia, the Department of Education carries out the administration of the tests with assistance from the principals of the schools and the Grades 3 and 5 class teachers.

The BSTP instruments consist of three major sections: (a) student questionnaire, (b) Numeracy test, and (c) Literacy test.

The same student questionnaire was used at Grades 3 and 5, and on all the six testing occasions in South Australia. This questionnaire contained items that asked students to provide information about some aspects of their home background. Consequently, the student questionnaire contained items that were used to construct student-level variables that could be used in the subsequent analyses in this study to identify student-level factors influencing student achievement in the Basic Skills Tests. However, the main BSTP data contained no information regarding the schools such as the school's absenteeism rate, mobility rate and locality. Since such school information is essential for a clear understanding of the factors involved in student achievement at the school-level, the information was obtained from the DETE in South Australia. Although the school data obtained from DETE lacked details regarding the students in the schools, it nevertheless contained vital information that was used to construct school-level variables that could be employed in the subsequent analyses in this study to identify school-level factors influencing student achievement in the Basic Skills Tests.

The 1994 to 2000 Numeracy tests at both grade levels consisted of items that covered three areas of numeracy, namely: Number, Measurement and Space. The 1994 to 2000

Literacy tests at both grade levels consisted of two sub-tests: (a) the Language sub-test, and (b) the Reading sub-test.

For a particular testing occasion, a few common items were included in each sub-test of the BSTs with the purpose of comparing performance between the Grade 3 and Grade 5 levels. However, there were no common items included from the previous testing occasions. Hence, in this study, the comparison of performance between testing occasions cannot be directly achieved by use of common item equating. However, because the Basic Skills Tests that are used in NSW are the same tests as used in South Australia, some equating data were obtained from NSW Department of School Education to help link the tests from the different occasions in South Australia. The NSW equating data consisted of groups of Grades 3 and 5 students from NSW who had taken the 1996 test (or another test directly linked to the 1996 test) as a trial test a week prior to taking the real test for that occasion. The procedure followed to equate the tests from the six different testing occasions is provided in Chapter 6 of this book.

By and large, the data available provide important school-level information and student-level information as well as measures in Numeracy and Literacy on two occasions (Grade 3 and Grade 5) for a substantial number (37,832) of the same students and across the different testing occasions (1995 to 2000) for a substantial number (426) of the same schools. Consequently, the data available provide strong possibilities of estimating value added measures of learning gain over the learning periods from Grade 3 to Grade 5 and for examining the stability of the value added measures across the different occasions and across the two subject areas of school learning, numeracy and literacy. This chapter leaves no doubts that there are sufficient data for meaningful analyses to be undertaken and reported in this study.

4

Methods

In order to answer the research questions raised in Chapter 1, it requires that several methods should be employed to analyze the data. Generally, the analyses in this study fall into two categories. First, there are those analyses that have to be carried out so as to develop common scales on which the achievement of the students across the two grade levels and across the six testing occasions can be assessed or compared. Second, there are those analyses that have to be carried out to identify the factors involved in students' achievement and to compute indices for assessing the performance of the primary schools in South Australia using the scores from the BST.

This chapter describes the methods of analyses under the second category. The first section of this chapter focuses on multilevel modelling by introducing the main computer package used to undertake the multilevel analyses in this study. The second section of this chapter describes the general method employed to estimate school effects in this study. Descriptions of the methods of analysis falling under the first category, that is, development of common scales, are presented in Chapter 6 together with the results of the analysis.

Multilevel modelling

The discussions in the current section focus mainly on the use of the HLM computer program in multilevel modeling especially in identification of the factors involved in students' achievement.

A hierarchical linear modelling technique is used in this study because this technique has strong literature backing (Chapter 2) for analysis of data collected in more than one level, as it is the case in this study (Chapter 3).

HLM computer program

The main computer package used for the multilevel analyses in this study is *HLM5 for Windows* developed by Raudenbush, Bryk and Congdon (2000). The HLM program was initially developed to find a solution for the methodological weakness of educational research during the 1980s, which was the failure of many quantitative studies to attend to the hierarchical, multilevel character of much educational research

data (Bryk and Raudenbush, 1987 & 1992). This failure came from the fact that "the traditional linear models used by most researchers require the assumption that subjects respond independently to educational programs" (Raudenbush and Bryk; 1994, p. 2590). In practice, most educational research studies select students as a sample who are nested within classrooms, and the classrooms are in turn nested within schools, schools within geographical location, state, or country. In this situation, the students selected in the study are not independent, but rather nested within organizational units and ignoring this fact results in the problems of "aggregation bias and misestimated precision" (Raudenbush and Bryk; 1994, p. 2590).

Among the programs included in this HLM5 package are HLM5/2L and HLM5/3L programs for statistical modelling of two- and three-level data structures, respectively. These two programs, that is, HLM/2L and HLM5/3L, are used in the current study. From the use of either of these two programs, effects of student-level factors and school-level factors can be examined at the same time. In addition, it is also possible to examine interaction effects of the factors at different levels on the outcome variables.

Specification of HLM model

Hierarchical linear modelling implies that regression equations are used to represent the model at each level of hierarchy. Consequently, Hox (1995; p. 11) notes that the full multilevel model "can be viewed as a hierarchical system of regression equations". As a result, the model specification in HLM analysis is undertaken by using the regression equation at each level.

For instance, achievement in numeracy at Grade 5 may be hypothesized to be influenced by gender of student (SEX) and age of student (AGE), location of the school (METRO), and size of the school (SSIZE). Within a two-level analysis, the regression equation at the student-level for this hypothesized model would be:

$$\mathbf{Y}_{ij} = \beta_{0j} + \beta_{1j}\mathbf{SEX}_{ij} + \beta_{2j}\mathbf{AGE}_{ij} + \mathbf{r}_{ij}$$

where:

$\mathbf{Y}_{ij}$ is the outcome variable (Y5NSCORE, numeracy score at Grade 5) of student *i* in school *j*;

β_{0j} is the intercept (that is, the mean achievement) of school *j*;

β_{1j} is the regression slope associated with SEX for school *j*;

β_{2j} is the regression slope associated with AGE for school *j*;

$\mathbf{r}_{ij}$ is a random error or 'student effect', that is, the deviation of the student mean from the school mean.

At the school-level, the regression equation for this hypothesized model would be:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\mathbf{METRO}_j + \gamma_{02}\mathbf{SSIZE}_j + \mathbf{u}_{0j}$$

where

γ_{00} is the expected intercept for the predictors METRO and SSIZE,

γ_{01} is the regression slope associated with METRO, the locality of the school,

γ_{02} is the regression associated with SSIZE, the size of the school, and

$\mathbf{u}_{0j}$ is a random effect or 'school effect' associated with school *j*.

If it is further hypothesized that there are interaction effects between METRO and SSIZE at the school-level with AGE at the student-level on the outcome variable (Y5NSCORE), then the regression equation for these interaction effects is:

$$\beta_{2j} = \gamma_{20} + \gamma_{21}\mathbf{METRO}_j + \gamma_{22}\mathbf{SSIZE}_j + \mathbf{u}_{2j}$$

where

β_{2j} is the regression slope associated with AGE for school *j*;

γ_{20} is the expected intercept for the predictors METRO and SSIZE for β_{2j};

γ_{21} is the regression slope associated with the interaction effect between METRO with AGE and the outcome variable Y5NSCORE;

γ_{22} is the regression slope associated with the interaction effect between SSIZE with AGE and the outcome variable Y5NSCORE; and

$\mathbf{u}_{2j}$ is a random effect associated with AGE.

The same approach described above for specifying two-level HLM models is used to specify three-level HLM models.

Running HLM

HLM analyses using either the HLM5/2L or the HLM5/3L program are undertaken by first constructing the sufficient statistics matrices (SSM) file, followed by the execution of analyses based on the SSM file (Bryk et al., 1996; p. 9). The creation of the SSM file involves the selection of the data and the variables to be included in the analyses at each level of the hierarchy.

After successful creation of the SSM file, the next step is to create a command file to execute the desired HLM analyses. At this stage, a decision is made on the model (equations) to be analyzed at each level of the hierarchy. This stage usually involves either three or four steps depending on whether it is a two-level or a three-level analysis being undertaken. The initial three steps are the same regardless of whether a two- or a three-level analysis is being undertaken.

The first step involves running a so-called 'null' model, which is also called a 'fully unconditional' model. The null model is the simplest model because it has no predictors at any level of the hierarchy. The purpose of running a null model is to estimate the amounts of variance in the outcome variable at the various levels of hierarchy (Bryk and Raudenbush, 1992) and also to test whether there are significant differences between the cluster units in the model with respect to the criterion variable.

The second step involves adding Level-1 predictors into the model, but without entering predictors at any of the other levels of the hierarchy. This model is called the 'unconditional' model at Level-1 and its purpose is to examine which Level-1 variables have significant effects on the outcome variable. The estimated coefficients (slopes) of each Level-1 predictor can either be ‘fixed’, which constrain the slopes and intercepts to be the same across all the Level-2 and/or Level-3 units, or ‘random’, which allows them to vary among Level-2 and/or Level-3 units (Raudenbush et al., 2000). The adding up of predictors can be undertaken either by a so-called 'step-down' procedure or a 'step-up' procedure. Under the step-down procedure, all Level-1 variables are entered into the equation simultaneously, and then the variables that do not have a significant influence on the outcome variables are deleted one at a time. Under the step-up procedure, the Level-1 variables are entered into the equation one at a time, consecutively, building up the equation by deleting any non-significant variables.

The third step, which is the final step for a two-level analysis, involves adding Level-2 predictors to the model. For a two-level analysis, this is the full model but for a three-level analysis, it is the so-called 'unconditional' model at Level-2. The listing of predictors at this step can be undertaken either by a step-down procedure or a step-up procedure described above. For a three-level analysis, the estimated coefficients of each Level-2 predictor could either be 'fixed', which constrain the slopes and intercepts to be the same across all the Level-3 units, or 'random', which allows them to vary among Level-3 units.

The fourth step, which is the final step for a three-level analysis, involves adding Level-3 predictors into the model using either the step-down procedure or a step-up procedure described above.

It should be noted that there are three options available for entering the predictors into the model under HLM5, regardless of whether the step-down or the step-up procedure is used. These three options are (a) adding a predictor 'uncentred' (b) adding a predictor 'group-mean centred' (b) adding a predictor 'grand-mean centred'. The second option 'group-mean centred' is not available at Level-3 of the analysis. The choice of the centering option to use depends on the nature of the predictor and the purpose of the research. Raudenbush et al. (2000) say that what is important is for the researcher to keep in mind the centering option used when interpreting the results. Kreft (1995) and Kreft et al. (1995) have provided details on the meaning of the results under different centering options.

Interpreting results

The output generated by the HLM5/2L and HLM5/3L computer programs provide information about reliability estimates, fixed effects, variance components and the deviance statistics. A discussion for each type of information is presented below.

Reliability estimates

The output generated by HLM5/2L and HLM5/3L provide reliability estimates at Level-1 of the model for each variable with random effects at the level. In addition, the output generated by HLM5/3L provides reliability estimates at Level-2 of the model for each variable with random effects at that level. These reliability estimates can simply be interpreted as indicators of the proportion of variance among the variable that can be considered true parameter variance. For example, a reliability estimate of 0.750, means that 75.0 per cent of the variance among the estimates of the variable can be considered true parameter variance; the remaining 25.0 per cent is random fluctuation that could be associated with measurement and sampling error (Bryk and Raudenbush, 1992). Indeed, the higher the reliability estimates of the parameters in the model, the better the model. Nevertheless, Bryk and Raudenbush, (1992) argue that it is generally possible to undertake HLM estimations with reliabilities as low as 0.05. Details on how the reliability estimates are calculated can be found in Bryk and Raudenbush (1992; p. 43).

Fixed effects

The fixed effects results in the output generated by HLM5/2L and HLM5/3L provides information about (a) the path coefficients and the standard error associated with each path coefficient, and (b) t-ratio and p-value, which indicate the significance of the path coefficient.

Path coefficients obtained using raw scores of the variables are normally called 'metric' or ‘unstandardized’ regression coefficients while those obtained using standardized scores of the variables are called 'standardized' regression coefficients. The sizes of standardized regression coefficients of the variables indicate the relative magnitude of effects and can therefore be used to rank the variables in terms of their relative degree of influence on the outcome within the same sample (Hox, 1995). However, the sizes of metric regression coefficients of the variables do not indicate the relative magnitude of effects and can not therefore be used to compare the degree of influence of the variables on the outcome. Nevertheless, the metric regression coefficients are useful where the aim of the analysis is to compare different samples to each other (see Hox, 1995; p.26).

In most studies, any variables with effects whose t-ratios are below |2.00| and whose p-values are larger than 0.05 are regarded as not significant and such variables are normally removed from subsequent analysis (For example, Lietz, 1996; Mohandas, 1999). For the current study, unless where it is otherwise stated, any effect with a t-ratio of below |2.00| and a p-value of larger than 0.05 is regarded as not significant and is removed from further analysis.

Deviance statistic

HLM5/2L and HLM5/3L compute a deviance statistic for the model tested plus the number of parameters in the model for each run. Hox (1995) and Raudenbush et al. (2000) have said that deviance statistics may be viewed as a measure of model fit because the higher the deviance, the poorer the fit of the model. Regarding this test, Hox points out that when one model is a subset of another model, the difference between their deviance is distributed as a chi-square, with degree of freedom equal to the difference of the number of parameters included in the two models. (Hox, 1995; p.34)

By a subset model, Hox refers to a model that, is a more complex model in that it includes all the parameters of another simpler model plus one or more additional parameters.

In addition, Raudenbush et al. (2000) have said that this chi-square (also known as a variance-covariance components) test is best used if the Full Maximum Likelihood (MLF) procedure is employed as the estimation mode and not when the Restricted Maximum Likelihood (MLR) estimation procedure is used. This is because under the MLR estimation procedure, the number of parameters remains the same between two models which differ only in their regression coefficients and therefore the chi-square test can only be used to examine the fit of the unconditional part of the model. Hox (1995) has suggested that the standard errors and the associated p-value for each coefficient could be used to check the fit of the model after the unconditional parts of the model are estimated using the MLR procedure.

For the purposes of comparing the fit of the models using the deviance statistics, an optional hypothesis testing sub-routine, that is available in HLM5/2L and HLM5/3L, is employed to compare model fit in successive HLM runs. This is done by entering the deviance statistic and number of parameters reported in the output file of a previous model into the optional hypothesis testing dialog box fields provided in the computer program. A chi-square statistic, with associated degrees of freedom and p-value are then printed at the end of the next HLM5/2L (or HLM5/3L) output file.

Variance components

Each run of HLM5/2L (or HLM5/3L) provides variance components for each level in the hierarchy. From these results of the variance components, the following can be calculated: (a) the proportions of variance available at each level; and (b) the proportion of the variance available at each level that is explained, and (c) the proportion of the total variance available that is explained. An outline of the calculations involved is given in the next sub-sections. A discussion of these calculations is to be found in Bryk and Raudenbush (1992; p. 60-76).

Variance partitioning

The proportion of variance available at each level is calculated from the results of variance components in the null model as follows.

$$\text{Prop. of var. available at the level} = \frac{\text{(Variance component at the level)}}{\text{(Total variance available)}} \qquad \text{Equation 4.1}$$

Thus, for a two-level analysis, the proportions of variance available at Levels 1 and 2 are obtained as follows:

$$\text{Proportion of variance available at Level-1} = \sigma_0^2 / (\sigma_0^2 + \tau_0) \qquad \text{Equation 4.2}$$

$$\text{Proportion of variance available at Level-2} = \tau_0 / (\sigma_0^2 + \tau_0) \qquad \text{Equation 4.3}$$

where:

σ_0^2 and τ_0 are the null two-level model variance components at Levels 1 and 2, respectively.

And for a three-level analysis, the proportions of variance available at Levels 1, 2 and 2 are obtained as follows:

$$\text{Proportion of variance at Level-1} = \sigma_0^2 / (\sigma_0^2 + \tau_{\pi 0} + \tau_{\beta 0}) \qquad \text{Equation 4.4}$$

$$\text{Proportion of variance at Level-2} = \tau_{\pi 0} / (\sigma_0^2 + \tau_{\pi 0} + \tau_{\beta 0}) \qquad \text{Equation 4.5}$$

$$\text{Proportion of variance at Level-3} = \tau_{\beta 0} / (\sigma_0^2 + \tau_{\pi 0} + \tau_{\beta 0}) \qquad \text{Equation 4.6}$$

where:

σ_0^2, $\tau_{\pi 0}$ and $\tau_{\beta 0}$ are the null three-level model variance components at levels 1, 2 and 3, respectively.

Variance explained at each level

The proportions of variance explained at each level of the final model are calculated from the variance component of the null model at the level and the variance components of the final model at the level as follows:

$$\text{Prop. of Var. Exp. at the level} = \frac{\text{(Null Var. Comp. at the level) - (Model's Variance Comp. at the level)}}{\text{(Null Var. Comp. at the level)}} \qquad \text{Equation 4.7}$$

Hence, for a two-level analysis, the proportions of variance explained at Levels 1 and 2 are obtained as follows:

$$\text{Proportion of variance explained for Level-1} = (\sigma_0^2 - \sigma_f^2) / (\sigma_0^2) \qquad \text{Equation 4.8}$$

$$\text{Proportion of variance explained for Level-2} = (\tau_0 - \tau_f) / (\tau_0) \qquad \text{Equation 4.9}$$

where:

σ_0^2 and τ_0 are the null two-level model variance components at Levels 1 and 2 respectively; and

σ_f^2 and τ_f are the final two-level model variance components at Levels 1 and 2 respectively.

And for a three-level analysis, the proportions of variance explained at Levels 1, 2 and 3 are obtained as follows:

$$\text{Proportion of variance explained for Level-1} = (\sigma_0^2 - \sigma_f^2) / (\sigma_0^2) \quad \text{Equation 4.10}$$

$$\text{Proportion of variance explained for Level-2} = (\tau_{\pi 0} - \tau_{\pi f}) / (\tau_{\pi 0}) \quad \text{Equation 4.11}$$

$$\text{Proportion of variance explained for Level-3} = (\tau_{\beta 0} - \tau_{\beta f}) / (\tau_{\beta 0}) \quad \text{Equation 4.12}$$

where:

σ_0^2, $\tau_{\pi 0}$ and $\tau_{\beta 0}$ are the null three-level model variance components at Levels 1, 2 and 3, respectively; and

σ_f^2, $\tau_{\pi f}$ and $\tau_{\beta f}$ are the final three-level model variance components at Levels 1, 2 and 3, respectively.

Total variance explained

The proportion of total variance explained at each level of the final model is calculated by multiplying the proportion of variance explained at the level with the proportion of variance available at the level, that is:

$$\text{Prop. of Total Var. Exp. at the level} = (\text{Prop. of Var. Exp. at the level}) \times (\text{Prop. of Var. available at the level}) \quad \text{Equation 4.13}$$

By substituting Equations 4.1 and 4.7 into the Equation 4.13, gives the following equations for total variance explained at levels 1 and 2 for a two-level analysis:

$$\text{Prop. of Total Var. Exp. at the Level-1} = \{(\sigma_0^2 - \sigma_f^2)/(\sigma_0^2)\} \times \{\sigma_0^2/(\sigma_0^2 + \tau_0)\} \quad \text{Equation 4.14}$$

$$\text{Prop. of Total Var. Exp. at the Level-2} = \{(\tau_0 - \tau_f)/(\tau_0)\} \times \{\tau_0/(\sigma_0^2 + \tau_0)\} \quad \text{Equation 4.15}$$

where:

σ_0^2 and τ_0 are the null two-level model variance components at Levels 1 and 2, respectively; and

σ_f^2, and τ_f are the final two-level model variance components at Levels 1 and 2, respectively.

Similarly, the equations for total variance explained at levels 1, 2 and 3 for a three-level analysis are:

$$\text{Prop. of Total Var. Exp. at the Level-1} = \{(\sigma_0^2 - \sigma_f^2) / (\sigma_0^2)\} \times \{\sigma_0^2 /(\sigma_0^2 + \tau_{\pi 0} + \tau_{\beta 0})\} \quad \text{Equation 4.16}$$

$$\text{Prop. of Total Var. Exp. at the Level-2} = \{(\tau_{\pi 0} - \tau_{\pi f}) / (\tau_{\pi 0})\} \times \{\tau_{\pi 0} /(\sigma_0^2 + \tau_{\pi 0} + \tau_{\beta 0})\} \quad \text{Equation 4.17}$$

$$\text{Prop. of Total Var. Exp. at the Level-3} = \{(\tau_{\beta 0} - \tau_{\beta f}) / (\tau_{\beta 0})\} \times \{\tau_{\beta 0} /(\sigma_0^2 + \tau_{\pi 0} + \tau_{\beta 0})\} \quad \text{Equation 4.18}$$

where:

σ_0^2, $\tau_{\pi 0}$ and $\tau_{\beta 0}$ are the null three-level model variance components at levels 1, 2 and 3, respectively; and

σ_f^2, $\tau_{\pi f}$ and $\tau_{\beta f}$ are the final three-level model variance components at levels 1, 2 and 3, respectively.

The proportion of total variance explained in the final model is then obtained by adding up the proportions of variance explained at each level of the final model and from the resulting figure, the proportion of variance left unexplained is computed by subtraction.

Residuals

The 'basic specification' option in HLM5/2L and HLM5/3L can be used to generate residual files that allow the examination of the structure of the residuals. HLM5/2L produces a Level-2 residual file while HLM5/3L produces two residual files, one at Level-2 and one at Level-3. These files contain the empirical Bayes (EB) residuals defined at Level-2 (and Level-3 for HLM5/3L), fitted values, and ordinary least (OL) residuals. These files can be exported into other data analysis programs (such as SPSS 10.0) for further analysis.

Bryk and Raudenbush (1992; pp.39-44; 76-82) and Raudenbush et al. (2000) have described how these residuals could be used to check the goodness of fit of the model. For example, in hierarchical linear modelling the EB residuals are assumed to be normally distributed and, therefore, it is important to check the adequacy of this assumption for significance testing. Histogram plots of the EB residuals for each of the predictor variables (and the intercepts) that have their random effect not fixed provide a convenient check for the adequacy of the assumption.

Apart from examination of goodness of fit of a model, the EB residuals could also be used to examine school effect, that is, the contribution of a school to the increase in student achievement. More details on the use of residuals in examination of school effects are provided in the next section and Chapters 9 to 11.

Estimation of school effects

This section describes the general model and concepts used to estimate school effects in this study.

Willms and Raudenbush (1989; pp.212-14) and Raudenbush and Willms (1995; pp.313-19) presented a two-level hierarchical linear model, which hypothesizes that a student's academic outcome (Y) is influenced by three general factors: the student background characteristics (S), school context (C), and school policies and practices (P). They illustrated how this model can be used to estimate Type A and Type B school effects. The theoretical and statistical concepts of this model by Willms and his colleague are adopted for use in this study because several recent studies have strongly commended the model for formulation of school effect indices (for example, Harker and Nash, 1996; Meyer, 1996 & 1997; Pituch, 1999 William et al., 2000; Ehrenberg et al., 2001). Furthermore, this model is derived from Carroll's model of school learning (Carroll, 1963), which many researchers accept as the most appropriate model of school learning currently available (see Creemers, Scheenens and Reynolds, 2001; pp.283-285).

The model by Willms and Raudenbush is described in some details below to help illustrate its application in the current study.

Following the above notations, the within-school regression model can simply be written:

$$\mathbf{Y}_{ij} = \boldsymbol{\beta}_{0j} + \mathbf{S}_{ij} + \mathbf{r}_{ij} \qquad \text{Equation 4.19}$$

where:

$\mathbf{Y}_{ij}$ is the outcome score for student *i* in school *j*;

$\boldsymbol{\beta}_{0j}$ is the mean achievement of school *j*;

$\boldsymbol{S}_{ij}$ is the contribution of the background characteristics of student *i* in school *j* (for example, gender, age and prior achievement) and;

$\mathbf{r}_{ij}$ is a random error or 'student effect', that is, the deviation of the student mean from the school mean.

The indices *i* and *j* denote students and schools where there are

$i = 1, 2, \ldots, n_j$ students within school j; and

$j = 1, 2, \ldots, J$ schools.

Willms and Raudenbush (1989) argue that if the student background characteristics (S) are grand-mean centred in HLM analysis, then the estimates of the intercepts, $\boldsymbol{\beta}_{0j}$, are the background-adjusted school means, and they describe how well a student with a sample-average background characteristic can be expected to score in each school.

The second level of the two-level model regresses the adjusted school performance, $\boldsymbol{\beta}_{0j}$, on the various school-level variables that describe school context (C) and school policy and practice (P):

$$\boldsymbol{\beta}_{0j} = \boldsymbol{\gamma}_{00} + \mathbf{C}_j + \mathbf{P}_j + \mathbf{u}_{0j} \qquad \text{Equation 4.20}$$

where:

$\boldsymbol{\gamma}_{00}$ is the grand mean,

$\mathbf{C}_j$ is the contribution of school context (for example, aspects of school composition such as the average socioeconomic level of the students in the school);

$\mathbf{P}_j$ is the effects of school policy and practice (for example, aspects of school leadership, use of resources, curricular content, and classroom instructional strategies);

$\mathbf{u}_{0j}$ is a school-level residual also called a random 'school effect', that is, the deviation of the school mean ($\boldsymbol{\beta}_{0j}$) from the grand mean ($\boldsymbol{\gamma}_{00}$) and it "represents the unique contribution of each school that is not explained by school-level variables in the model" (Willms and Raudenbush, 1989; p. 212).

Equations 4.19 and 4.20 can be combined into a single equation to yield the following model, which describes the linear relationship of the components involved.

$$\mathbf{Y}_{ij} = \boldsymbol{\gamma}_{00} + \mathbf{C}_j + \mathbf{P}_j + \mathbf{u}_{0j} + \mathbf{S}_{ij} + \mathbf{r}_{ij} \quad \text{or}$$

$\mathbf{Y}_{ij} = \boldsymbol{\gamma}_{00}$ ..... (grand mean)

$+ \mathbf{S}_{ij}$ .....(contribution of student background characteristics)

$+ \mathbf{C}_j + \mathbf{P}_j + \mathbf{u}_{0j}$ ..... (contribution of school context, policy and practice)

$+ \mathbf{r}_{ij}$ ..... (student-level random error)

Equation 4.21

Based on the above model (Equation 4.21), Willms and Raudenbush (1995) have defined two types of school effects. The first is Type A, defined as:

$$\mathbf{A}_j = \mathbf{C}_j + \mathbf{P}_j + \mathbf{u}_{0j} \qquad \text{Equation 4.22}$$

Thus, Type A effect includes the effects of school context, policy and practice, and therefore, it is an indicator of how well a student of average background characteristics would perform in school *j*, relative to the grand mean. Consequently, Raudenbush and Willms (1995; p.310) argue that the Type A effect "is the effect parents generally consider when choosing one of the *J* schools for their child". They further argue that it would clearly be unfair to reward those who work in the school on the basis of Type A effects because the school staff is only partly responsible for those effects.

The second is the Type B effect, defined as:

$$\mathbf{B}_j = \mathbf{P}_j + \mathbf{u}_{0j} \qquad \text{Equation 4.23}$$

Thus, the Type B effect includes only the effects of school policy and practice, and therefore, it is an indicator of how well a particular school performed relative to other schools with similar student intake and context. It is important to note that Type B effect excludes factors that lie outside the control of those who work in the school. Consequently, Raudenbush and Willms (1995; p.310) argue that Type B effect "is the effect school officials consider when evaluating the performance of those who work in the schools".

Raudenbush and Willms (1995) report two strategies for estimating school effects based on the model specified above (Equation 4.21); namely, the 'addition' approach and the 'subtraction' approach.

In order to estimate school effects using the addition approach, all the relevant variables (that is, student background characteristics, school context, policy and practice) are identified and measured so that the model given by Equation 4.21 is fully specified. If the model given by Equation 4.21 is fully specified it follows that Type A and Type B effects can be estimated by addition using Equations 4.22 and 4.23, respectively.

In order to estimate Type A effects using the subtraction approach, the model given by Equation 4.21 is estimated with measures of student background characteristics but without data describing school context and policies. The estimates of $\mathbf{u}_{0j}$ (school-level residual) are then the estimates of Type A effects. Similarly, to estimate Type B effects under the subtraction approach, the model given by Equation 4.21 is estimated with measures of student background characteristics and school context but without data describing school policies. The residual term ($\mathbf{u}_{0j}$) then includes the effects of school policy, practices and other unmeasured effects.

Raudenbush and Willms (1995) argue that the estimation of school effects under the subtraction approach can only be achieved without bias if the school context and policies are orthogonal. However, they have illustrated that the bias issue is not a major problem if the estimation of school effects is undertaken through the empirical Bayes procedure (see Raudenbush and Willms, 1995; p.328-330); the estimation procedure employed in HLM5 computer program.

In this study, the relevant data for school policy and practices are not available, and therefore, the subtraction approach is used to estimate both types of school effects. More detailed description of the actual models used to estimate the various types of school effects based on the general model (Equation 4.21) and concepts described above are presented in Chapters 9 to 11 together with the results of the analyses.

Summary

This chapter provide descriptions of the methods of data analysis and the computer program selected to carry out the necessary multilevel analyses to (a) identify the factors involved in student achievement, and (b) to compute indices for assessing the performance of the primary schools in South Australia. The analyses of data in this study had to take into consideration the multilevel nature of the data. HLM5 is the one program that is readily available and which takes into account the multilevel nature of the data.

Chapters 7 and 8 present more detailed description of the analyses carried out to identify the factors involved in student achievement using two-level and three-level models, respectively, while Chapters 9 to 11 present descriptions of the analyses carried out to estimate school effects.

5

Design and Models

The issues addressed in this study fall in to three broad categories. First, there are those issues concerned with the investigation of the achievement levels of the Grades 3 and 5 students in the Basic Skills Tests in South Australia. Second, there are issues concerned with the investigation of the factors influencing student achievement of numeracy and literacy. Third, there are issues concerned with formulating indices for assessing the performance of the primary schools in South Australia using the scores from the BST.

Consequently, this chapter describes the general design and models employed in this study to address the above issues. The design and the models were derived from the research questions presented in Chapter 1 and were restricted by the problems in the data involved as discussed in Chapters 1 and 3.

The first section of this chapter describes the general design employed to link data sets from the six testing occasions in South Australia and outlines the propositions advanced in this study that deal with levels of achievement of the Grades 3 and 5 students in numeracy and literacy. The second section describes the general models used in this study to tease out factors influencing student achievement in the Basic Skills Tests while the third section describes the general model used to estimate indices of school performance.

Design employed to link data sets

In this study, Grade 3 and Grade 5 data within the same testing occasion are linked by common items while data across the testing occasions are linked by common persons. Common persons are used to link the data across the testing occasions because there are no items included on more than one testing occasion in the BSTP. In South Australia (SA), however, there was no common persons information collected to link the data from the different testing occasion.

Conveniently, the Basic Skills Tests that are used in New South Wales (NSW) are the same as the ones used in SA, and the Department of Schools Education in NSW collects information on common persons for purposes of linking data from the different testing occasions in that State. Consequently, data on common persons were

obtained from the NSW Department of School Education and these data are used to link the SA data from the different testing occasions. The data from NSW consists of groups of Grades 3 and 5 students from that State who had taken the 1996 test (or another test directly linked to the 1996 test) as a trial test a week prior to taking the real test for that occasion.

The data sets involved in this study have already been described in Chapter 3. Figure 5.1 is a diagram matrix representation of the general design employed to link SA data sets from the six testing occasions. In interpreting the diagram shown in Figure 5.1, it should be borne in mind that this study is designed with replication across two curriculum areas, that is, literacy and numeracy.

The shaded rectangles appearing under the South Australia Data section of the diagram (Figure 5.1) represent the Grade 3 and Grade 5 data sets from each of the six testing occasions (1995 to 2000) as shown in the diagram. In this connection, the overlaps appearing in the diagram between the Grade 3 and Grade 5 data sets are used to show that the SA data from the same testing occasion are brought together through the use of common items. On the other hand, the shaded rectangles appearing under the New South Wales Data section of the diagram represent the common persons' data sets from NSW used to join the SA data sets from the six testing occasions. In the diagram, the pattern of the shading used for a particular data set in the NSW section is such that it corresponds tothe pattern of shading for the SA data sets linked by that data set. For example, the pattern used in the rectangle on the top left hand corner under the 1994 NSW data section is the same pattern as that used in the rectangles representing 1995 SA data sets. For this example, it means that the 1995 data are linked to the 1994 data through a group of students from NSW who took the 1994 test as a trial test a week prior to taking the real test for the 1995 testing occasion. It should be noted, however, that for reasons of parsimony, not all the tests taken by the common persons from NSW are represented in the diagram in Figure 5.1. From the diagram in Figure 5.1, it should also be noted that the data sets from SA are linked back to the 1996 SA data sets through baseline data sets from the NSW either directly or indirectly. For example, the diagram shows that the 1995 SA data sets are linked to 1994 NSW baseline data set, which is in turn linked to the 1996 NSW data set, and consequently, linked to 1996 SA data sets. Similarly, the 1998 and 2000 SA data sets are linked to the 1996 SA data set through the 1997 and 1999 NSW baseline data sets, respectively.

In addition, the diagram in Figure 5.1 identifies the four cohorts of students from SA involved in this study. For example, Cohort 1 consists of the students who took the test at Grade 3 in 1995 and at Grade 5 in 1997, Cohort 2 consists of students who took the test at Grade 3 in 1996 and at Grade 5 in 1998, and so on. In order to reflect the year of participation by the cohort at Grades 3 and 5, some sections of this report refer to Cohort 1 as the 1995/1997 cohort, Cohort 2 as the 1996/1998 cohort, Cohort 3 as the 1997/1999 cohort, and Cohort 4 as the 1998/2000 cohort.

In summary, the diagram in Figure 5.1 shows how the NSW baseline data sets joins the SA data sets from all the six testing occasions and identifies the four cohorts of students involved in this study. In modelling with the school as the main unit of interest, the diagram in Figure 5.1 shows that there are either six data points or four data points for the schools depending on whether the six testing occasions or the four cohorts are considered as the observation points.

The design shown in Figure 5.1 makes it possible to compare the performance of the students in the Basic Skills Tests across Grades 3 and 5 levels and across the six testing occasions or across the four cohorts of students involved in this study.

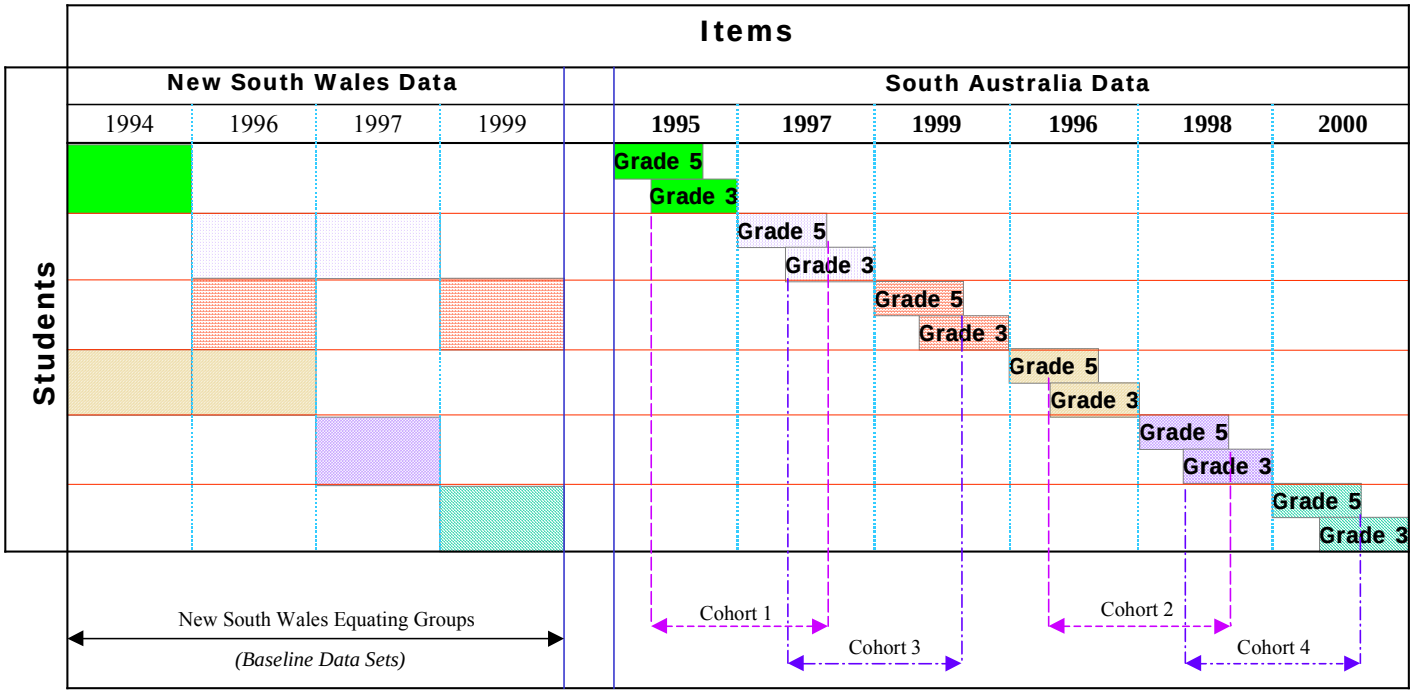


Figure 5.1 Design employed to link data sets within and across occasions

Thus, with the above design (Figure 5.1) and test equating techniques described in Chapter 6, it is possible to test propositions directed towards the research questions (presented in Chapter 1) that deal with the levels of achievement of the Grades 3 and 5 students in numeracy and literacy. These research questions are listed below.

1. Is there adequate fit of the Rasch model to the Grades 3 and 5 items?
2. How do the average item difficulties of the Grades 3 and 5 tests compare across testing occasions?
3. Can the numeracy items for 1995 to 2000 tests for Grades 3 and 5 be brought to a common scale?
4. Can the literacy items for 1995 to 2000 tests for Grades 3 and 5 be brought to a common scale?

Hypothesized models for achievement factors

It should be borne in mind that the data sets for this study were collected at different levels. When dealing with multilevel data such as the data in this study, the appropriate procedure is to formulate multilevel models, "which enable the testing of hypotheses about effects occurring within each level and the interrelations among them" (Raudenbush and Bryk, 1994; p. 2590). The formulation of multilevel models is justified because in data of a multilevel nature, the effects of the variables under investigation operate within each level and are interactive across levels. Thus, for this study, it is necessary to formulate multilevel models to enable testing of hypotheses about the factors influencing student achievement in numeracy and literacy using the HLM5 computer package (Raudenbush, Bryk and Congdon, 2000).

With the multilevel nature of the data in mind (that is, student, school and occasion), two general types of multilevel models were formulated for teasing out factors influencing student achievement in this study: two-level models and three-level models. For each of the general type of model, three models were formulated for testing. Descriptions of these models as well as the details of the data sets involved in each of these models are provided in the next two sub-sections, with the first sub-section dealing with the two-level models and the second sub-section dealing with the three-level models. For reasons of parsimony, the discussions in the sections that follow do not include descriptions of any hypothesized cross-level interaction effects. However, it should be noted that, in the actual analyses several cross-level interaction effects are examined.

Two-level models

Figures 5.2, 5.3 and 5.4 show the three two-level hierarchical linear models formulated for testing in this study. The hierarchical structure of all the two-level models formulated for testing in this study is such that students are nested within schools, that is, the Level-1 units are students and Level-2 units are schools.

In this study, the three models shown in Figures 5.2, 5.3 and 5.4 are simply referred to as *Model-X, Model-Y* and *Model-Z* respectively. For each of the three types of models proposed, two separate models are specified, one for Numeracy and the other for Literacy. The names and definitions of the variables tested for inclusion in each level of these two-level models are provided in Table 5.1. Details on the formation and the coding of these variables have already been provided in Chapter 3.

Model-X

The outcome variables of interest in Model-X (Figure 5.2) are scores for the Numeracy (NSCORE) and Literacy (LSCORE) tests at Grades 3 and 5 levels. At the student-level, it is hypothesized that six independent variables directly influence student achievement in numeracy (or in literacy). The six independent variables are: YEARLEVL, SEX, AGE, ATSI, HOME (or NESB) and INOZ.

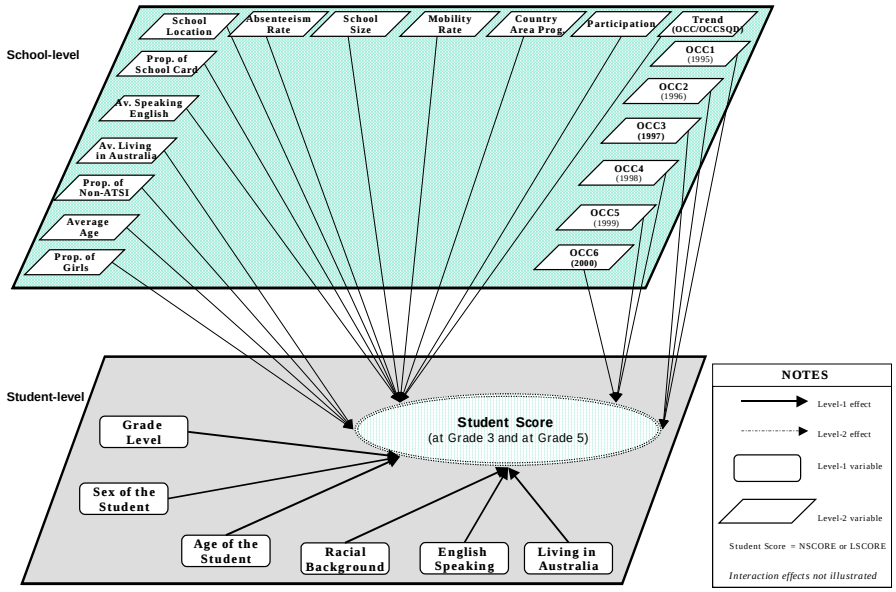


Figure 5.2 Two-level hierarchical model for numeracy and literacy - Model-X

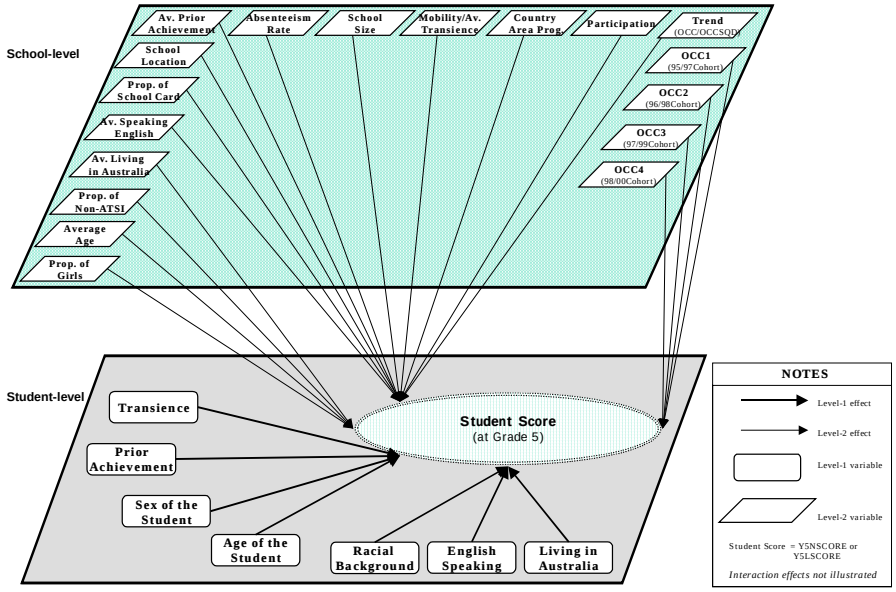


Figure 5.3 Two-level hierarchical model for numeracy and literacy - Model-Y

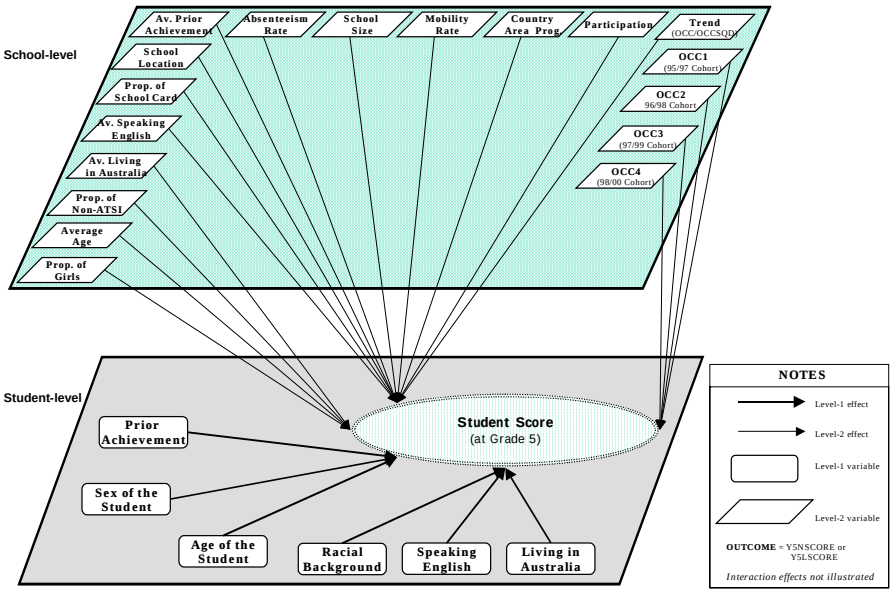


Figure 5.4 Two-level hierarchical model for numeracy and literacy - Model-Z

In addition, it is hypothesized that 19 variables at the school-level also directly influence student achievement in numeracy (or literacy). Of these 19 variables at the school-level, five are aggregated from student-level variables (SEX_1, AGE_1, ATSI_1, INOZ_1, HOME_1/NESB_1), six are school information variables obtained from DETE files (PSCARD, METRO/GPOLOG, ABSENT, SSIZELOG, MOBILITY, CAP), six are occasion dummy variables (OCC1 to OCC6), two are trend variables (OCC or OCCSQD) and one is a school participation variable (YR35PPT).

Generally, Model-X is aimed at teasing out the factors influencing student achievement across the Grades 3 and 5 levels by analysing all the BSTP data collected from the six testing occasions of interest in this study. Consequently, the model involves all the 106,514 students and all the 504 schools that have taken part in the BSTP from 1995 to 2000 (see Chapter 3).

However, the maximum possible number of student-level units in Model-X is 144,346 because of the 106,514 students 37,832 have data at Grade 3 and at Grade 5, while 22,348 have data only at Grade 5, and 25,227 have data only at Grade 3. In addition, in the data there are 294 and 257 missing scores (either Grade 3 or Grade 5) for numeracy and literacy respectively because some students did not respond to at least one item in tests and therefore their scores were not estimated. However, in total, only four students have missing scores at both grade levels. Because these amounts of missing data are not substantial compared to the number of cases in the data, the pairwise option for deletion of cases with missing data was selected in the construction of the HLM Sufficient Statistics Matrix (SSM) for Model-X. The pairwise deletion process left a total of 144,342 Level-1 units and led to the elimination of a total of three Level-2 units that had inadequate data.

Table 5.1 Variables tested in the two-level models

	Variable label	Variable Name (Description)
Level-1 (Student)		
	AGE	Age of the Student
	ATSI	Racial Background (Aboriginal or Torrens Strait Islanders)
	HOME	Speaking English at Home
	INOZ	Living in Australia
	LSCORE	Literacy Score (Grade 3 or Grade 5 Literacy Score)
	NESB	Non-English Speaking Background
	NSCORE	Numeracy Score (Grade 3 or Grade 5 Numeracy Score)
	SEX	Sex of the Student
	TRANS	Transience
	Y3LSCORE	Prior Achievement (Grade 3 Literacy Score)
	Y3NSCORE	Prior Achievement (Grade 3 Numeracy Score)
	Y5LSCORE	Literacy Score (Grade 5 Literacy Score)
	Y5NSCORE	Numeracy Score (Grade 5 Numeracy Score)
	YEARLEVL	Grade Level
Level-2 (School)		
	ABSENT	Absenteeism Rate in the School
	AGE_1	Average Age of the Students in the School
	ATSI_1	Proportion of Non-ATSI Students in the School
	CAP	Country Area Program
	GPODIST	School Location (Dist. of the School from Adelaide GPO
	GPOLOG	School Location (Logarithm of GPODIST)
	HOME_1	Average Speaking English at Home
	INOZ_1	Average Living in Australia
	METRO	School Location (Rural/Urban)
	MOBILITY	Mobility Rate of the School
	NESB_1	Proportion of English Speaking Students in the School
	OCC	Trend (Occasion, Linear Trend)
	OCC1	1995 in Model-X, 1995/1997 Cohort in Models Y and Z
	OCC2	1996 in Model-X, 1996/1998 Cohort in Models Y and Z
	OCC3	1997 in Model-X, 1997/1999 Cohort in Models Y and Z
	OCC4	1998 in Model-X, 1998/2000 Cohort in Models Y and Z
	OCC5	1999 in Model-X
	OCC6	2000 in Model-X
	OCCSQD	Trend (Occasion Squared, Quadratic Trend)
	PSCARD	Proportion of School Cardholders in the School
	SEX_1	Proportion of Girls in the School
	SSIZE	School Size (Number of Students in the School)
	SSIZELOG	School Size (Logarithm of SSIZE)
	TRANS_1	Average Transience (Prop. of Newcomers at Grade 5)
	Y3LSCO_1	Average Prior Achievement (Av. Grade 3 Literacy Score)
	Y3NSCO_1	Average Prior Achievement (Av. Grade 3 Numeracy Score)
	YR35PPT	Participation Size of the School in the BSTP

The maximum possible number of school-level units in Model-X is 2,871 because using HLM5/2L a unique identity had to be employed for each Level-2 unit, and therefore, different identities were used to represent each school on the various testing occasions. Furthermore, out of the 504 schools only 489 participated in 1995, 485 in 1996, 482 in 1997, 474 in 1998, 473 in 1999, and 468 in 2000, which made the maximum possible number of school-level units to be 2,871. However, as mentioned above, three Level-2 units were dropped using the pairwise deletion of cases with missing data, which left 2,868 Level-2 units in the SSM file for Model-X.

In Model-X, the year of study (YEARLEVL) is included as a predictor and also to differentiate between the Grade 3 and the Grade 5 scores, and hence, prior achievement of the student is not available for examination in this model. Nonetheless, Model-X provides a plausible approach to teasing out the factors influencing student achievement in the BST especially if the data that are available are from one testing occasion only. Moreover, because Model-X includes the year of study as a predictor, the model could be appropriate in the estimation of growth in achievement across the two grade levels in South Australia and in each school included in the analyses.

Model-Y

The outcome variables of interest in Model-Y (Figure 5.3) are scores for the Numeracy and Literacy tests at Grade 5, represented by Y5NSCORE and Y5LSCORE respectively. At the student-level, it is hypothesized that seven independent variables directly influence student achievement in numeracy (or in literacy); namely, Y3NSCORE (or Y3LSCORE), TRANS, SEX, AGE, ATSI, HOME (or NESB) and INOZ. It is also hypothesized that 20 variables at the school-level also directly influence student achievement in numeracy (or literacy). Apart from Y3NSCO_1 (or Y3LSCO_1) and TRANS_1, all the school-level variables tested in this model are the same as the ones tested in Model-X.

Generally, Model-Y is aimed at teasing out the factors influencing student achievement by analysing the so-called ‘transience data set’ from the BSTP. The transience data set involves all the students who could be matched regardless of whether the student remained in the same school or whether the student changed schools between the two grades. Thus, the model involves all the 37,832 students who have data both at Grade 3 and at Grade 5 levels. 32,741 of the students included in this model had remained in the same schools they attended in Grade 3, while 5,091 had changed schools. However, 71 and 61 students in the data have no Grade 3 scores in numeracy and literacy respectively, while 54 and 39 students in the data have no Grade 5 scores in numeracy and literacy respectively. As a result, pairwise deletion of cases with missing data employed in the construction of the SSM file for Model-Y, leaves a total of 37,824 Level-1 units as a result of the elimination of five Level-2 units from the analyses because they have inadequate data. Although only 482 schools are included in Model-Y, the total number of Level-2 units in this model is 1,853 because using HLM5/2L unique identity has to be employed for each Level-2 unit in the analyses.

In general, Australian society is highly mobile (Fields, 1995), and therefore, student mobility is inevitable in South Australia. In Model-Y, the variable TRANS (transience) is included to differentiate the Grade 5 students who changed schools from the Grade 5 students who remained in the same school they attended in Grade 3. Thus, the model could be employed to estimate the effects of transience on student achievement across the two grade levels in South Australia. Moreover, apart from teasing out the factors influencing student achievement, Model-Y also provides a plausible approach for assessing the performance of schools in South Australia in

terms of the value added to student achievement over the two-year period. However, for Model-Y to be employed in assessment of performance of the schools, it has to be assumed that it is appropriate to award the value added component to the Grade 5 school the student is in and disregards any contribution that might have been made to the component by the school where the student was in Grade 3 or possibly Grade 4.

Model-Z

In Model-Z, the outcome variables of interest are also scores for the Numeracy and Literacy tests at Grade 5, represented by Y5NSCORE and Y5LSCORE, respectively (Figure 5.3). Model-Z differs from Model-Y only because it aims at teasing out the factors influencing student achievement by analysing the so-called 'non-transience data set' from the BSTP. The non-transience data set involves the students who could be matched and at Grade 5 they were in the same schools they had attended in Grade 3.

Thus, Model-Z involves only the 32,741 students who could be matched in the same schools. For this model, 55 and 43 students in the data had no Grade 3 scores in numeracy and literacy respectively, while 40 and 29 students in the data had no Grade 5 scores in numeracy and literacy respectively. The pairwise deletion of cases with missing data employed in the construction of the SSM file for Model-Z, lelf a total of 32,732 Level-1 units and as a result of the elimination of six Level-2 units from the analyses because they had inadequate data. Although only 480 schools are included in Model-Z, the total number of Level-2 units in this model is 1,823 because using HLM5/2L unique identity has to be employed for each Level-2 unit in the analyses.

Apart from teasing out the factors influencing student achievement, Model-Z also provides the approach that is most fair for assessing school performance in the State in terms of the value added to student achievement over the two-year period because all the contributions made to student achievement can be directly attributed to a particular school.

Three-level models

It has been mentioned above that, using HLM5/2L, unique identity has to be employed for each Level-2 unit in the analyses. Consequently, in the two-level models described in the previous sub-section, each school is treated as a different school on each testing occasion. Because of the multilevel nature employed by the HLM5/2L computer program, at the planning stages of this study there were concerns about the appropriateness of using the two-level models described in the previous sub-section to tease out the factors influencing student achievement. In addition, there were concerns about the appropriateness of partitioning of variance and monitoring of linear (or quadratic) trends in achievement using the two-level models given that the multilevel nature employed in HLM5/2L computer program does not link data of the same school from the different testing occasions. Therefore, a decision was made to reformulate the three models described above in terms of a three-level structure and, accordingly, employ the HLM5/3L (Raudenbush et al., 2000) computer program. The multilevel model employed in HLM5/3L allows the identity of the school to be kept intact over time.

Figures 5.5, 5.6 and 5.7 show the three three-level hierarchical linear models formulated for teasing out factors influencing student achievement in this study and referred to as Model-X, Model-Y and Model-Z respectively. The three-level models formulated here correspond directly to the two-level models described in the previous sub-section. However, it should be noted that the hierarchical structure of these three-

level models is such that students are nested within school, and schools in turn are nested within occasions, that is, the Level-1 units are students, Level-2 units are schools and Level-3 units are occasions. It should also be borne in mind that for each of the three types of models formulated, two separate models are specified, one for Numeracy and the other for Literacy. The names and definitions of the variables tested for inclusion in each level of these three-level models are provided in Table 5.2. Details on the formation and the coding of these variables are provided in Chapter 3.

At the student-level, the structures of the three-level models are exactly the same as the structures of the two-level models. At the school-level, unlike the two-level models, three-level models do not include the dummy variables denoting the testing occasions or the student cohorts (OCC1 to OCC6) and the trend variables (OCC and OCCSQD), and instead these variables are included in a level of their own, the occasion-level or macro-level, a third level.

Thus, Model-X involves all the students who have taken part in the BSTP from 1995 to 2000, while Model-Y involves only those students who could be matched (transience data set), and Model-Z involves only those students who could be matched and had remained in the same schools over the two-year period (non-transience data set). Employing the pairwise deletion of cases with missing data in the construction of the SSM files provides the number of units at Levels 1, 2 and 3 respectively as follows: 144 342, 2 868 and 6 for Model-X; 37 824, 1 853 and 4 for Model-Y; and 32 732, 1 823 and 4 for Model-Z. The numbers of units at Levels 1 and 2 in the three-level models are exactly the same as the number of units obtained at the levels in the corresponding two-level models described in the previous sub-section.

Unlike the two-level models described in the previous sub-section, the three-level models described in this sub-section allow the identity of the school to be kept intact over time in HLM analyses. In addition, HLM analyses of the three-level models enable the amounts of variances available and explained at the occasion-level to be disentangled from the amounts of variances available and explained at the school-level, giving a better image of the whole system.

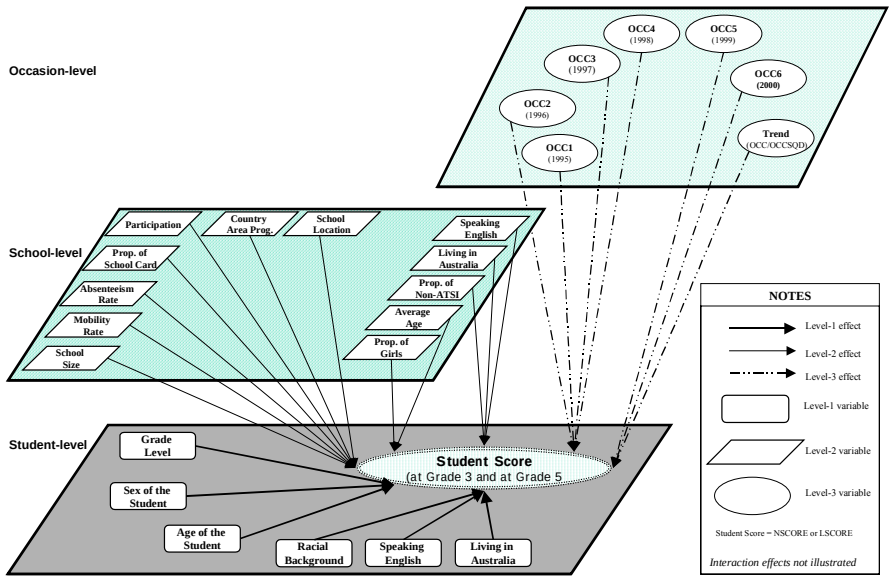


Figure 5.5 Three-level hierarchical model for numeracy and literacy – Model-X

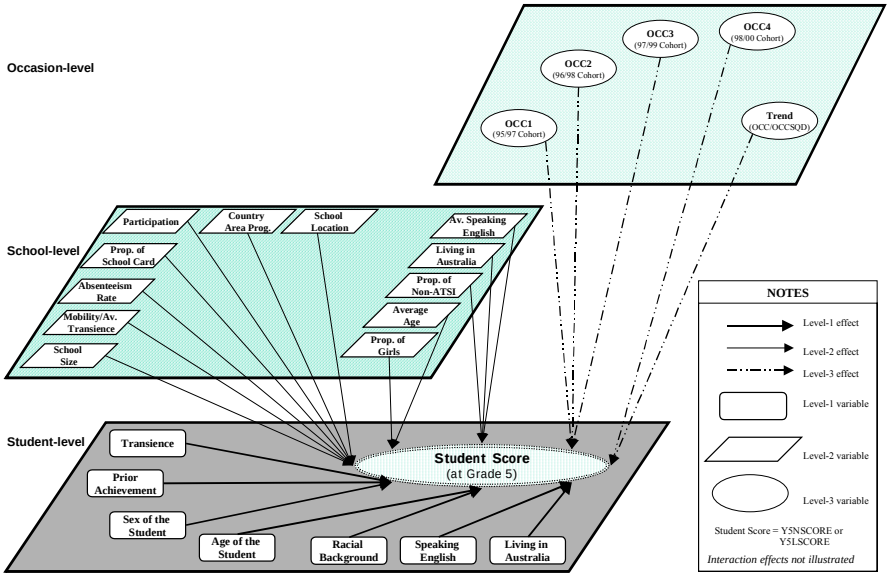


Figure 5.6 Three-level hierarchical model for numeracy and literacy – Model-Y

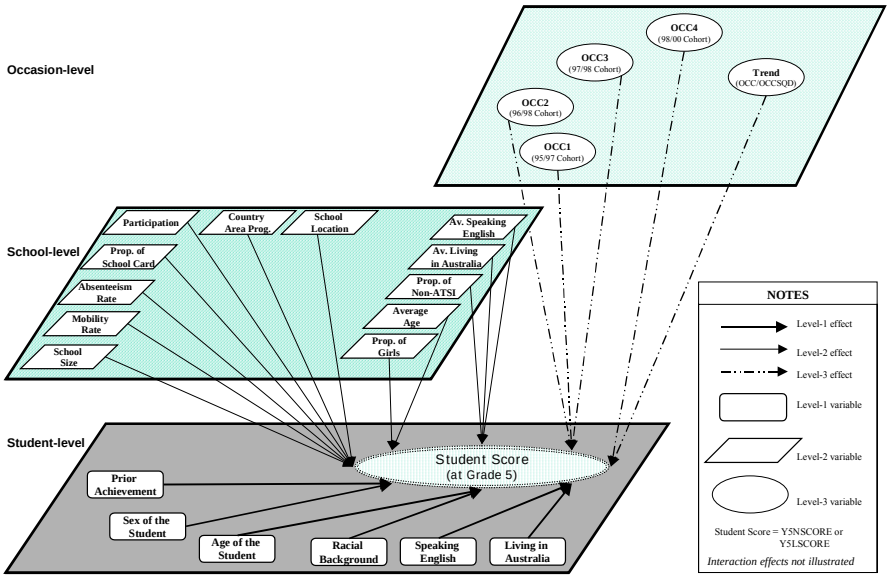


Figure 5.7 Three-level hierarchical model for numeracy and literacy – Model-Z

Nonetheless, there were concerns regarding the appropriateness of the three-level models in significance testing especially at the third level (occasion-level) where the numbers of units are small, that is, six (testing occasions) for Model-X and four (student cohorts) for Models Y and Z. Consequently, for purposes of comparing results from the two types of models, a decision was made to carry out HLM analyses on both the two-level (Chapter 7) and three-level (Chapter 8) models described above.

Table 5.2 Variables tested in the three-level models

	Variable label	Variable Name (Description)
Level-1 (Student)		
	AGE	Age of the Student
	ATSI	Racial Background (Aboriginal or Torrens Strait Islanders)
	HOME	Speaking English at Home
	INOZ	Living in Australia
	LSCORE	Literacy Score (Grade 3 or Grade 5 Literacy Score)
	NESB	Non-English Speaking Background
	NSCORE	Numeracy Score (Grade 3 or Grade 5 Numeracy Score)
	SEX	Sex of the Student
	TRANS	Transience
	Y3LSCORE	Prior Achievement (Grade 3 Literacy Score)
	Y3NSCORE	Prior Achievement (Grade 3 Numeracy Score)
	Y5LSCORE	Literacy Score (Grade 5 Literacy Score)
	Y5NSCORE	Numeracy Score (Grade 5 Numeracy Score)
	YEARLEVL	Grade Level
Level-2 (School)		
	ABSENT	Absenteeism Rate in the School
	AGE_1	Average Age of the Students in the School
	ATSI_1	Proportion of Non-ATSI Students in the School
	CAP	Country Area Program
	GPODIST	School Location (Distance of the School from Adelaide GPO)
	GPOLOG	School Location (Logarithm of GPODIST)
	HOME_1	Average Speaking English at Home
	INOZ_1	Average Living in Australia
	METRO	School Location (Rural/Urban)
	MOBILITY	Mobility Rate of the School
	NESB_1	Proportion of English Speaking Students in the School
	PSCARD	Proportion of School Cardholders in the School
	SEX_1	Proportion of Girls in the School
	SSIZE	School Size (Number of Students in the School)
	SSIZELOG	School Size (Logarithm of SSIZE)
	TRANS_1	Average Transience (Prop. of Newcomers at Grade 5
	Y3LSCO_1	Average Prior Achievement (Av. Grade 3 Literacy Score)
	Y3NSCO_1	Average Prior Achievement (Av. Grade 3 Numeracy Score)
	YR35PPT	Participation Size of the School in the BSTP
Level-3 (Occasion)		
	OCC	Trend (Occasion, Linear Trend)
	OCC1	1995 in Model-X, 1995/1997 Cohort in Models Y and Z
	OCC2	1996 in Model-X, 1996/1998 Cohort in Models Y and Z
	OCC3	1997 in Model-X, 1997/1999 Cohort in Models Y and Z
	OCC4	1998 in Model-X, 1998/2000 Cohort in Models Y and Z
	OCC5	1999 in Model-X
	OCC6	2000 in Model-X
	OCCSQD	Trend (Occasion Squared, Quadratic Trend)

With the two-level and three-level models described above, it is possible to test propositions directed towards the research questions (presented in Chapter 1) that dealt with the factors influencing student achievement in numeracy and literacy. These research questions are listed below.

5. Has the level of performance in numeracy (or literacy) at Grade 5 changed significantly over time?
6. What is the average growth in numeracy and literacy achievement between Grades 3 and 5 levels?
7. What student-level factors influence numeracy (or literacy) achievement?
8. What school-level factors influence numeracy (or literacy) achievement?
9. What cross-level interaction effects influence numeracy (or literacy) achievement?
10. What amounts of variance are available at the student-level, school-level and occasion-level?
11. What percentages of variance in student scores in numeracy and literacy at Grade 5 are explained by Prior Achievement (that is, achievement at Grade 3) alone?
12. What percentages of variance in student scores in numeracy and literacy at Grade 5 do the predictor variables in the final two-level and three-level models explain?

Hypothesized models for estimation of school effects

Apart from teasing out factors influencing student achievement, the two- and three-level models that involve students with two data points (that is, Models Y and Z) described in the preceding sections could also be employed to estimate the school's contribution to student achievement across Grades 3 and 5. However, two problems are easily recognisable if the models described above were to be employed to estimate indices of individual school effectiveness.

First, if the two-level models described above were to be employed to estimate the indices, there would be a different index for the school for each cohort of students involved. The idea of averaging these different indices to obtain an overall index for each school is not appealing because this averaging would not take into consideration the multilevel nature of the data and, therefore, the resulting overall index would suffer from design effect problems. Second, if the three-level models described above were to be employed to estimate the indices, the number of units at the third level is too small (four cohorts) for significance testing at that level. Moreover, a third problem arises because under the multilevel structure employed in the two- and three-level models described above, the factors that influence the performance slope can not be investigated.

Because of the problems mentioned above, a decision was made to reformulate Models Y and Z based on a structure that that would allow (a) the estimation of one overall index for each school while at the same time taking into consideration the multilevel nature of the data involved, and (b) for the factors influencing a performance slope to be investigated.

Thus, the models shown in Figures 5.8 and 5.9 were reformulated for estimating indices of individual school effectiveness and investigating the factors influencing performance slope in this study. The model shown in Figure 5.8 uses the transience data set (that is, data on all the students who could be matched, N=37,832), and

therefore it is a transience model. The model shown in Figure 5.9 use the non-transience data set (that is, data on the students matched in the same school, N=32,741), and therefore it is a non-transience model. In other words, the data sets analysed in the models shown in Figures 5.8 and 5.9 are the same data sets analysed in Models Y and Z, respectively.

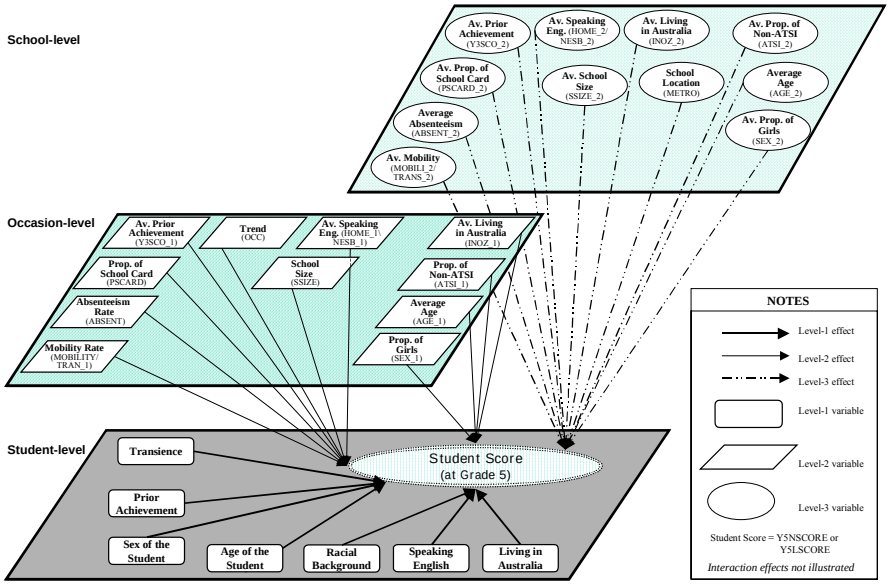


Figure 5.8 Model for estimation of school effects using the transience data set

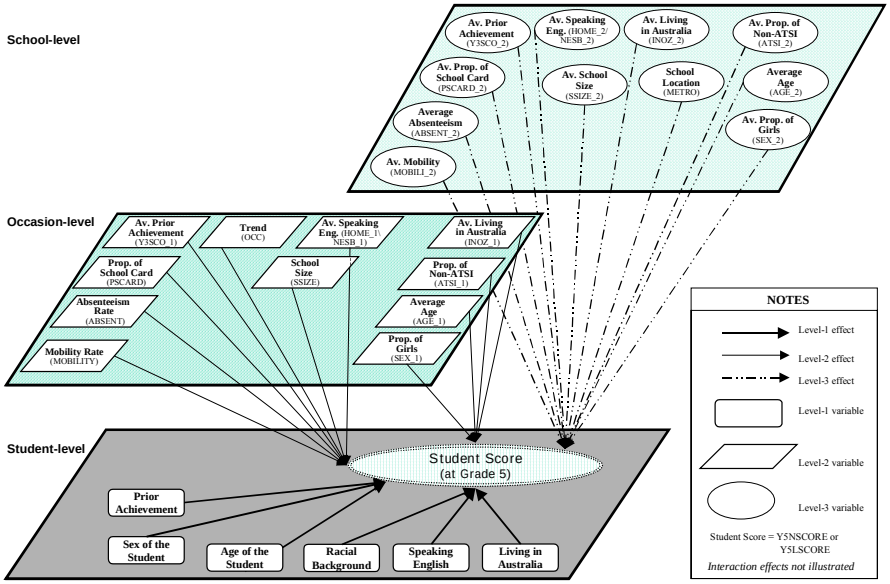


Figure 5.9 Model for estimation of school effects using the non-transience data set

For each of the models shown in Figures 5.8 and 5.9, two separate models are specified, one for Numeracy and the other for Literacy. The names and definitions of the variables tested for inclusion in each level of these models are provided in Table 5.3. Generally, in these models, the variables tested for inclusion at the second-level are formed by aggregating student-level data from only one testing occasion while variables tested for inclusion at the third-level are formed by aggregating student-level data from all the four testing occasions. More detailed descriptions on how these variables were formed and coded were provided in Chapter 3.

It should be noted that although a three-level structure is employed in the models shown in Figures 5.8 and 5.9, their hierarchical nature are different from the three-level model described in the preceding section. The Level-1, Level-2 and Level-3 for the models shown in Figures 5.8 and 5.9 are student, occasion and school whereas for the three-levels model described in the preceding section these levels are student, school and occasion respectively. It should further be noted that, under the multilevel structure employed in the models shown in Figures 5.8 and 5.9, data on successive cohorts of students are used to map performance of the schools. Hence, for the schools, the structure employed in these two models comprises a longitudinal or multiwave design (Bryk and Raudenbush, 1992). Importantly, apart from using this multiwave design to estimate one overall index of performance for each school, the design coupled with the hierarchical linear modelling technique forms a powerful approach for measuring change in the performance of individual schools over time.

The above approach is powerful because it overcomes two serious limitations of conventional analytical procedures. First, the approach overcomes the well-documented difficulties encountered in the measurement of change rooted in the misconception that argues that change should be viewed as an increment, as the difference between before and after estimates of performance (Willett 1988), that is, a pre-post or two-wave design. In a multiwave design, individual change is followed over time at sensibly spaced intervals and if the growth is changing steadily and smoothly over time, three or four spaced measurements on each individual may capture the shape and direction of the change. Second, because this approach incorporates hierarchical linear modelling techniques, the approach solves the problem often encountered in conventional analytical procedures, that is, the inability to distinguish differences in rates of change among individuals schools (Bryk and Raudenbush, 1992; Raudenbush and Bryk, 2002).

Willms and Raudenbush (1989) employed a similar longitudinal structure to estimate school effects and their stability using data from Scotland that consisted of representative samples of two cohorts of students who completed their secondary education in 1980 (N=1,500) and 1984 (N=5,000) respectively. Because the current study analyses data from four cohorts of students and larger numbers of students per cohort when compared with the study by Willms and Raudenbush, it should provide a rigorous test regarding the appropriateness of employing this longitudinal structure in the estimation of school effects and their stability. Most important, unlike the study by Willms and Raudenbush, this study moves beyond the limitations of the pre-post design in the measurement of change.

Studies of school effects generally agree that the most practical and realistic approach to assessing school performance in terms of the value added to student achievement is to analyse data on only those students who remain in the same school over the study period. Data on students who remain in the same school are appealing to researchers because all the contribution made to an increase in student achievement can be directly attributed to one school and, therefore, provide a picture that is fairest assessment of the performance of the school.

Table 5.3 Variables tested in the three-level longitudinal model

Variable label	Variable Name (Description)
Level-1 (Student)	
AGE	Age of the Student
ATSI	Racial Background (Aboriginal or Torrens Strait Islanders)
HOME	Speaking English at Home
INOZ	Living in Australia
LSCORE	Literacy Score (Grade 3 or Grade 5 Literacy Score)
NESB	Non-English Speaking Background
NSCORE	Numeracy Score (Grade 3 or Grade 5 Numeracy Score)
SEX	Sex of the Student
TRANS	Transience
Y3LSCORE	Prior Achievement (Grade 3 Literacy Score)
Y3NSCORE	Prior Achievement (Grade 3 Numeracy Score)
Y5LSCORE	Literacy Score (Grade 5 Literacy Score)
Y5NSCORE	Numeracy Score (Grade 5 Numeracy Score)
Level-2 (Occasion)	
OCC	Trend (Occasion, Linear Trend)
ABSENT	Absenteeism Rate of the School within the Occasion
AGE_1	Average Age of the Students in the School within the Occasion
ATSI_1	Proportion of Non-ATSI Students in the School within the Occasion
HOME_1	Average Speaking English at Home within the Occasion
INOZ_1	Average Duration of Living in Australia within the Occasion
MOBILITY	Mobility Rate of the School within the Occasion
NESB_1	Proportion of English Speaking Students in the School within the Occasion
PSCARD	Proportion of School Cardholders in the School within the Occasion
SEX_1	Proportion of Girls in the School within the Occasion
SSIZE	School Size (Number of Students in the School within the Occasion)
TRANS_1	Average Transience (Proportion of Newcomers at Grade 5 within the Occasion)
Y3LSCO_1	Average Prior Achievement (Av. Grade 3 Literacy Score within the Occasion)
Y3NSCO_1	Average Prior Achievement (Av. Grade 3 Numeracy Score within the Occasion)
Level-3 (School)	
ABSENT_2	Average Absenteeism Rate of the School over the Study Period
AGE_2	Average Age of the Students in the School over the Study Period
ATSI_2	Average Proportion of Non-ATSI Students in the School over the Study Period
HOME_2	Average Speaking English at Home over the Study Period
INOZ_2	Average of Living in Australia over the Study Period
METRO	School Location (Rural/Urban)
MOBILI_2	Average Mobility Rate of the School over the Study Period
NESB_2	Average Prop. of English Speaking Students in the School over the Study Period
PSCARD_2	Average Proportion of School Cardholders in the School over the Study Period
SEX_2	Average Proportion of Girls in the School over the Study Period
SSIZE_2	Average School Size (Av. Number of Students in the School over the Study Period)
TRANS_2	Average Transience (Av. Prop. of Newcomers at Grade 5 over the Study Period)
Y3LSCO_2	Average Prior Achievement (Av. Grade 3 Literacy Score over the Study Period)
Y3NSCO_2	Average Prior Achievement (Av. Grade 3 Numeracy Score over the Study Period)

In this study, however, both data on all the students who could be matched (thus, Figure 5.8), and on only those students who were matched in the same school (thus,

Figure 5.9), are analysed in order to examine whether or not the ranking orders of the schools obtained using the two data sets differ markedly.

Where the data consisting of all the students who could be matched are analysed (Figure 5.8), the model assumes that it is appropriate to assess the performance of the school in terms of the contribution made to student achievement, regardless of whether or not the student changes schools. In addition, the model assumes that it is appropriate to award the contribution made by the school to the increase in student achievement to the Grade 5 school the student is in, and disregards any contribution that might have been made by the school that the student was in at Grade 3 or Grade 4. Thus, if it is borne in mind that the Australian society is highly mobile (Fields, 1995), the transience model (Figure 5.8) could provide a plausible approach for assessing school performance in the State in terms of value added to student achievement over the two-year period.

It is worth noting that at Levels 1 and 2, the variables tested for inclusion in the longitudinal models described here (Figures 5.8 and 5.9) are mostly the same variables tested for inclusion in the corresponding three-level models described in the previous section. At the third level, however, the models described in this section for the estimation of school effects do not include the dummy variables denoting the occasions (OCC1 to OCC4) and the quadratic trend variable (OCCSQD), as is provided by the corresponding models described in the previous section. The reason for the exclusion of these occasion-related variables is apparent if it is remembered that, in estimation of school effects, researchers are mostly interested in variables that capture the school environment. Nevertheless, in order to examine changes in the performances of the schools, the linear trend variable (OCC) is included in the longitudinal models at Level-2. For the current study, this variable (OCC) is preferred over the quadratic trend variable (OCCSQD) because results from preliminary analyses indicate that a better fit of the models to the data is obtained when the former variable is employed.

It should be noted that, unlike studies that aim to tease out factors influencing student achievement where any variable can be included in the analysis, variables examined for inclusion in school effectiveness models must have some substantial theoretical backing. School effects studies have been subject to criticism for what is called 'fishing' for significant results without a theoretical basis or set of hypotheses to explain the results (see Coe and Fitz-Gibbon, 1998, p.424; Teddlie et al., 2001; p.73). It should be borne in mind that the general theoretical model for estimation of school effects, which was proposed by Willms and Raudenbush in 1989 and again in 1995, is employed in this study. The model proposed by Willms and Raudenbush states that student academic performance is influenced by three broadly defined factors: student background, school context, and school policies and practices (see Chapter 4). This model is derived from extensions of Carroll's model of school learning (Carroll, 1963).

Employing the pairwise deletion of cases with missing data in the construction of the SSM files provides the numbers of units at Levels 1, 2 and 3 respectively as follows: 37 824, 1 853 and 482 for the model shown in Figure 5.8; and 32 732, 1 823 and 479 for the model shown in Figure 5.9. The numbers of units at Levels 1 and 2 in the longitudinal models described here are exactly the same as the numbers of units obtained at the levels in the corresponding three-level models described in the previous section.

With the three-level longitudinal models described above, it is possible to test propositions directed towards the research questions (presented in Chapter 1) that deal

with the assessment of the performance of the primary schools in South Australia using the scores from the BST. These research questions are given below.

13. What percentages of variance in student scores in numeracy and literacy at Grade 5 are explained in the models employed to estimate school effects?
14. What percentages of variance in student scores in numeracy and literacy at Grade 5 are left unexplained at the student-level in the models employed to estimate school effects?
15. What percentages of variance in student scores in numeracy and literacy at Grade 5 are left unexplained at the school-level in the models employed to estimate school effects?
16. How reliably are the school effects estimated?
17. Can a stability index be calculated to compare the stability of the various types of school effects over time?
18. Based on value added scores, is the rank order of the schools, using all the students who could be matched, greatly different from the rank order of the schools using only those students who could be matched in the same school?
19. Are schools that are identified as relatively effective based on one type of school effect also identified as relatively effective based on a different type of school effect?
20. Do schools that show more than expected average levels of performance also show more than expected increases in performance over time?
21. Are schools that are relatively effective in numeracy also relatively effective in literacy?
22. Are schools that are relatively effective for one cohort of students also relatively effective for other cohorts of students?
23. Are schools that are relatively effective in numeracy for boys also relatively effective for girls?
24. Are schools that are relatively effective in literacy for girls also relatively effective for boys?

Several varieties of the longitudinal models described above (Figures 5.8 and 5.9) are tested in this study in the analyses undertaken to answer the above research questions. Analyses based on these models are reported in Chapters 9 to 11.

Summary

The design and models described in this chapter together with the data analysis methods described in Chapter 4 are necessary if answers to the research questions raised in Chapter 1 are to be obtained. The formulations of the design and the models are based on information that could be obtained from various data sets used in the study. Subsequent chapters discuss the analyses as well as the results of the analyses carried out based on the design and models described in this chapter and using the methods of analysis identified in Chapter 4.

6

Calibration and Equating

It has been noted in the introductory chapters that the Department of Education Training and Employment (DETE) of South Australia has administered the BST to Grades 3 and 5 students in government schools six times since the inception of the program in 1995. It has also been noted that the BSTP instruments consist of two major sub-tests: (a) the Numeracy test, and (b) the Literacy test. The Literacy test consists of two sub-tests: Language and Reading. The Numeracy test consists of items that cover three areas: Number, Measurement and Space. This chapter describes steps followed to equate all the Grade 3 and Grade 5 tests from the six occasions (1995 to 2000) to construct common scales: one for numeracy and the other for literacy. The common scales developed in this chapter are used in the construction of achievement related variables, (that is, student scores at Grades 3 and 5 in the Basic Skills Tests), which are used in subsequent analyses in this study.

This chapter first outlines the Rasch analyses of the tests, then describes vertical equating of the Grades 3 and 5 tests within the same occasion and, finally, the overall equating of the tests within and across occasions. It should be noted that this chapter provides only an outline of the steps followed to develop the common scales and to estimate the Rasch scores of the students for numeracy and literacy. Most of the theoretical details of the Rasch model as well employment of the model in scaling, equating and scoring of the test data can be found in Hungi (2003; pp.449-479). Detailed examples of how the QUEST (Adams and Khoo, 1999) computer program is used to carry out the analyses described in this chapter can also be found in Hungi (2003).

Rasch analyses

In this study, the items from all the six Numeracy tests as well as those from the six Literacy tests are examined for their fit to the Rasch model, which is the most robust of the Item Response Theory (IRT) models (Hambleton and Swaminathan, 1985; Skaggs and Lissitz, 1986). The aim of the Rasch analysis is to ascertain whether there are some items that need to be dropped from the tests when calibrating, equating and scoring.

The Infit Mean Square (INMS) is the item statistic used to detect misfitting items using the Rasch model. The INMS is an indication of how well the slope of the Item Characteristic Curve (ICC) of a given item fits that of the ideal (or expected) ICC at the threshold performance level (Wright and Stone, 1979; Wright and Masters, 1982). Adams and Khoo (1993) have suggested that an acceptable range for INMS would be from 0.77 to 1.30. Values below 0.77 indicate that the item discriminates too sharply between those candidates who are competent and those who are not competent. Hence, the item provides redundant information and has insufficient bandwidth. On the other hand, values beyond 1.30 indicate that the item has poor discrimination in that some relatively high performing students are getting the item wrong while some relatively low performing students are getting it correct.

Separate Rasch analyses were carried out on all the Grades 3 and 5 Basic Skills Test items using the QUEST (Adams and Khoo, 1999) computer program. The analyses were undertaken using all the cases that participated in the BST in South Australia on each of the six occasions and taking omitted and not-reached items as wrong. Any item falling outside INMS range of 0.77 to 1.30 was deleted.

All the items in the Grades 3 and 5 Numeracy tests for all the six occasions had their INMS values within the 0.77 to 1.30 range. However, two items for the Literacy tests fell beyond the stipulated INMS value (1.30) and were accordingly deleted (Wright and Stone, 1979; Smith and Kramer, 1992). The two items were Item 51 (INMS=1.31) in the 1997 Grade 5 and Item 13 (INMS=1.41) in the 2000 Grade 3 test.

From the Rasch analyses of the tests, it was evident that the vast majority of the items in the 1995 to 2000 BSTP had a high degree of consistency in terms of INMS values. The 1995 to 2000 tests had between them a total of 1,338 items (that is, 486 and 852 items for numeracy and literacy respectively) and only two were identified as misfitting. Hence, it was safe to conclude that the items in the BSTP were well constructed and developed because they had adequate fit to the Rasch model as indicated by their INMS statistics.

Equating within the same occasion

Equating within the same testing occasion was achieved through the use of the common items included by the BST developers in the Grade 3 and the Grade 5 tests. The number of common items included in the Numeracy and the Literacy tests are given in Table 6.1. The table also gives the percentage of the common items in each test as well as the total number of items in the combined Grades 3 and 5 Numeracy and also Literacy tests in the same testing year. Also presented in Table 6.1 are the total numbers of students who have participated in the BSTP in South Australia in each year level since 1995[13].

There are no globally acceptable numbers of common items needed to place two different test forms on one scale. However, there seems to be a general agreement among researchers that a better estimate of the common scale would be obtained if there were many common items. Nevertheless, some (for example, Smith and Kramer, 1992) argue that even as few as a single item could be employed to link two different test forms. Wright and Stone (1979) suggest that 10 to 20 items are needed to form a link between two different test forms consisting of 60 items each. This number is approximately 17 to 34 per cent of the items in each test. Hambleton et al. (1991)

[13] The statistics for 1994 test have been provided in Table 6.1 because they are referred to later in this chapter.

propose that the number of common items needs to be approximately between 20 and 25 per cent of the number of items in the tests.

In this study, all percentages of common items in the numeracy tests and the literacy tests for both year levels were within or above the ranges proposed in the wider literature (see Table 6.1). Hence, the numbers of common items were considered sufficient to link the Grade 3 and the Grade 5 tests within the same testing occasion.

The concurrent equating technique was applied to equate the Grade 3 tests to the Grade 5 tests within the same testing occasion. The concurrent method was chosen because, unlike the anchor and the common item differences equating techniques, it allowed for items that behaved differently from the other items in the combined data to be identified and removed from the analysis. Moreover, a number of studies have indicated that the concurrent technique yielded more consistent equating results (Kenyon and Stansfield, 1992; Shen, 1993; Mohandas, 1996).

Table 6.1 Numbers of Cases and Items in the SA BSTP data

			Numeracy				Literacy			
Year	Grade	Cases in SA	Items	Common Items	Total Items	% Common Items	Items	Common Items	Total Items	% Common Items
1994	Grade 3	None	32	9	67	28.1	59	15	125	25.4
	Grade 5	None	44			20.5	81			18.5
1995	Grade 3	10,283	32	9	71	28.1	57	22	114	38.6
	Grade 5	10,735	48			18.8	79			27.8
1996	Grade 3	11,095	32	10	70	31.3	59	20	119	33.9
	Grade 5	11,613	48			20.8	80			25.0
1997	Grade 3	12,437	32	11	69	34.4	58	21	120	36.2
	Grade 5	11,973	48			22.9	83			25.3
1998	Grade 3	12,794	32	10	70	31.3	61	20	124	32.8
	Grade 5	12,471	48			20.8	83			24.1
1999	Grade 3	12,550	35	11	72	31.4	63	20	127	31.7
	Grade 5	12,900	48			22.9	84			23.8
2000	Grade 3	12,677	35	12	71	34.3	62	21	124	33.9
	Grade 5	12,818	48			25.0	83			25.3
Total		**144,346**	**562**	**72**	**490**		**992**	**139**	**853**	

In order to equate the data concurrently for the Grades 3 and 5 tests within the same occasion, the two data sets from each occasion were merged to form one data set by bringing the common items in the two tests into the same columns. The calibration of the two tests was then done simultaneously counting omitted and not-reached items as wrong.

All the items in the combined Grades 3 and 5 Numeracy as well as in the Literacy tests had INMS values within the 0.77 to 1.30 range showing they still had adequate fit to the Rasch model after vertical equating.

Next, the thresholds of the 1995 to 2000 vertically equated test items were anchored and used to calculate the scores of all the students at the Grade 3 as well as at the Grade 5 levels counting omitted items and not-reached items as wrong using the Quest computer program (Adams and Khoo, 1999). The scores of the students with perfect scores and those of students with zero scores were estimated using the procedures described by Hungi (2003; pp. 475-9). A working rule was set not to calculate scores

for students who did not respond to at least one item in the tests. This scoring procedure was consistent with the approach employed by Australian Council for Educational Research (ACER) as well as the Basic Skills Tests developers in the department of Education in New South Wales.

The results of the vertical equating are presented in Table 6.2. The table gives the case mean and the item mean for numeracy and literacy for both year levels and for each of the six testing occasions. The differences between the Grades 3 and 5 case and item means for each of the six testing occasions are also provided in the table.

From the differences marked as 'A' and 'C' in Table 6.2, it can be observed that the level of performance of the Grade 5 students in either numeracy or literacy exceeded that of the Grade 3 students by approximately one logit on each testing occasion. The exceptions here were literacy 1997 and again 1998. The observed increase in performance supports the growth of approximately half a logit per year reported by Hungi (1997).

Nevertheless, for numeracy it should also be noted that there were slight but steady increases in the performance difference between the two grades from 1995 to 1999 and then a slight drop in 2000 (Table 6.2, difference 'A'). Similar observations are evident for literacy with increases in the performance increase recorded from 1995 to 1998, then a marked drop in 1999 and a slight recovery in 2000 (Table 6.2, difference 'C').

Table 6.2 Vertical equating results using SA data

a) Numeracy	1995	1996	1997	1998	1999	2000
i) Case Mean						
Grade 5	1.79	1.19	1.20	1.40	1.30	1.28
Grade 3	0.88	0.26	0.10	0.28	0.14	0.19
Difference (A)	0.91	0.93	1.10	1.12	1.16	1.09
ii) Item Mean						
Grade 5	0.26	0.48	0.45	0.39	0.33	0.31
Grade 3	-0.45	-0.96	-0.70	-0.65	-0.62	-0.58
Difference (B)	0.71	1.44	1.15	1.04	0.95	0.89

b) Literacy	1995	1996	1997	1998	1999	2000
i) Case Mean						
Grade 5	1.82	1.54	1.74	1.57	1.15	1.10
Grade 3	0.84	0.36	0.48	0.29	0.22	-0.01
Difference (C)	0.98	1.18	1.26	1.28	0.93	1.11
ii) Item Mean						
Grade 5	0.23	0.23	0.29	0.38	0.21	0.21
Grade 3	-0.62	-0.52	-0.55	-0.62	-0.28	-0.48
Difference (D)	0.85	0.75	0.84	1.00	0.49	0.69

The differences in the mean difficulty of the Grades 3 and 5 tests have fluctuated between 0.71-1.44 logits for numeracy and 0.49-1.00 logits for literacy (Table 6.2, differences 'B' and 'D'). These fluctuations are around 0.50 logits and account for about a year of learning for each subject (Hungi, 1997). Nevertheless, it should be appreciated that it is no easy task to develop and to select items of comparable level of

difficulty year after year especially given that the test papers of the previous occasions are left to circulate freely in the general community. Obviously, with just a few exceptions, the BST developers have generally maintained good judgment regarding the level of difficulty of the items to include at both year levels.

Equating across occasions

In the BSTP there are no items included on more than one testing occasion. Hence, comparison of performance between testing occasions could not be directly achieved in this study by the use of common item equating.

However, in New South Wales (NSW), all the Basic Skills Tests administered in that State were linked back to the 1996 test either directly (or indirectly through another occasion test) so as to equate tests across occasions (horizontal equating). Since the Basic Skills Tests that were used in NSW were the same ones used in South Australia (SA), some equating data were obtained from NSW Department of School Education to help link the tests from the different testing occasions in South Australia. These data consisted of groups of Grades 3 and 5 students from NSW who had taken the 1996 test (or another test directly linked to the 1996 test) as a trial test a week prior to taking the real test for that occasion. Figure 6.1 shows the overall equating design used to link the tests for all the six occasions. This diagram (Figure 6.1) is a simplified version of the diagram presented in Figure 5.1 in the previous chapter.

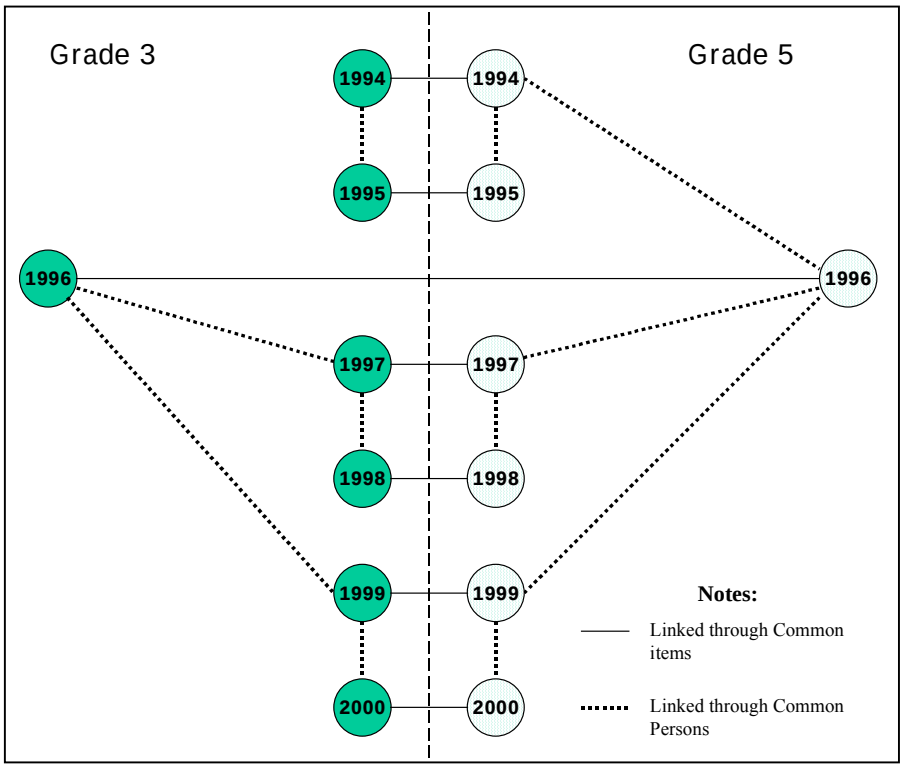


Figure 6.1 Overall equating design

From Figure 6.1, it can be noted that the 1995, 1998 and 2000 tests were not linked directly to the 1996 test but through the 1994, 1997 and 1999 tests respectively. It can also be noted that there was no direct link for the 1994 Grade 3 test to the 1996 Grade 3 test, and this was because no data were collected for that link in NSW. It should also be remembered that there were no data collected for 1994 in SA because the BSTP had not yet been introduced.

There were some concerns about the appropriateness of the equating design presented in Figure 6.1. In particular, there were some doubts concerning the double links involved in connecting the 1995, 1998 and 2000 tests to the 1996 tests. It was thought that the double links might distort the overall equating results to some extent.

It was also feared that lack of the 1994 NSW Grade 3 link data could distort the equating results to some extent. Furthermore, some analysis undertaken to examine the Grades 3 and 5 item means after vertical equating using NSW equating data produced results that differed substantially from the item means obtained when equating using SA students for both numeracy and literacy. Table 6.3 presents the Grades 3 and 5 item means obtained using NSW equating data and those obtained using SA students for both subject areas. The table also gives the total number of (Grade 3 plus Grade 5) students used for the vertical equating from each State.

The deviations given in Table 6.3 were obtained by subtracting the differences between the item means for Grade 3 and the item means for Grade 5 obtained using NSW equating groups (DNSW) from the corresponding differences obtained using SA cases (DSA).

Within the IRT framework (and when there is a close fit between the chosen IRT model and the test data set of interest), it would be expected that the item parameters (and therefore the item mean) would generally remain the same regardless of the sample of students used from the population of students for which the test was designed (Hambleton and Swaminathan, 1985; Skaggs and Lissitz, 1986; Petersen et al. 1989). However, in real test situations it is hard to get two samples of students who fit the chosen IRT model equally and therefore the item parameters obtained using any two samples of students can rarely be totally identical.

Nevertheless, the items means would be expected to remain fairly constant regardless of the student sample used since any little deviations in the item parameters would occur in both positive and negative directions and would therefore most likely cancel each other out. Consequently, very little (less than 0.13 logits[14]) or no differences were expected between the items means obtained using the NSW equating group and those obtained using SA cases. It was expected that any little differences that might be observed would be in both directions (to indicate random effects) and would most likely cancel out each other in the calculation of the deviation.

However, from Table 6.3 it can be observed that some of the deviations could be considered to have been substantially large (≥0.13 logits). For numeracy, two of the deviations (-0.30, and -0.13 for 1997, 1998 respectively) fell outside the set criteria compared to only one deviation (-0.14 for 2000) for the case of literacy. Perhaps the generally larger deviations observed for numeracy compared to literacy indicated greater differences in the numeracy syllabi compared to the literacy syllabi of the two States.

[14]The average growth in Literacy (or Numeracy) achievement between Grades 3 and 5 has been estimated as about 0.50 logits per year (Hungi, 1997). Hence, a deviation of 0.13 logits would indicate that the mean difficulty of the items differed by approximately one school-term's work and, therefore, it is substantial.

Table 6.3 Comparison of item means obtained using SA cases and using NSW equating groups

1) Numeracy

State		1995	1996	1997	1998	1999	2000
South Australia (all cases)	N_{SA}	21,018	22,708	24,410	25,265	25,450	25,495
	Grade 5 (A)	0.26	0.48	0.45	0.39	0.33	0.31
	Grade 3 (B)	-0.45	-0.96	-0.70	-0.65	-0.62	-0.58
	D_{SA} {=A - B}	0.71	1.44	1.15	1.04	0.95	0.89
New South Wales (equating group)	N_{NSW}	1,646	4,919	4,327	2,358	4,188	2,214
	Grade 5 (C)	0.27	0.51	0.55	0.45	0.41	0.33
	Grade 3 (D)	-0.46	-0.99	-0.90	-0.72	-0.66	-0.55
	D_{NSW} {=(C) - (D)}	0.73	1.50	1.45	1.17	1.07	0.88
	Deviation {=(D_{SA}) - (D_{NSW})}	**-0.02**	**-0.06**	**-0.30**	**-0.13**	**-0.12**	**0.01**

1) Literacy

State		1995	1996	1997	1998	1999	2000
South Australia (all cases)	N_{SA}	21,018	22,708	24,410	25,265	25,450	25,495
	Grade 5 (E)	0.23	0.23	0.29	0.38	0.21	0.21
	Grade 3 (F)	-0.62	-0.52	-0.55	-0.62	-0.28	-0.48
	D_{SA} {=E - F}	0.85	0.75	0.84	1.00	0.49	0.69
New South Wales (equating group)	N_{NSW}	1,646	4,919	4,327	2,358	4,188	2,214
	Grade 5 (G)	0.23	0.24	0.31	0.39	0.21	0.37
	Grade 3 (H)	-0.67	-0.53	-0.60	-0.61	-0.26	-0.46
	D_{NSW} {=G - H}	0.90	0.77	0.91	1.00	0.47	0.83
	Deviation {=(D_{SA}) - (D_{NSW})}	**-0.05**	**-0.02**	**-0.07**	**0.00**	**0.02**	**-0.14**

Notes:
D_{SA} - The Difference between the Grade 3 item mean and the Grade 5 item mean using SA cases.
D_{NSW} - The Difference between the Grade 3 item mean and the Grade 5 item mean using NSW equating cases.
N_{SA} - Total Number of Grade 3 plus Grade 5 cases in the SA data
N_{NSW} - Total Number of Grade 3 plus Grade 5 cases in the NSW equating data

The most interesting observation from Table 6.3 is the fact that almost all the deviations were negative regardless of the subject area indicating that the differences between the Grades 3 and 5 item means were greater when computed using the NSW equating sample than when computed using SA cases. In addition, a comparison of the Grade 5 tests items for numeracy obtained using SA cases (row 'A') with the corresponding item means obtained using NSW equating groups (row 'C') showed that the tests appeared harder when taken by the NSW equating students than when taken by the SA students. Similar observations were evident for the Grade 5 literacy item means (rows 'E' and 'G') though not as clear. However, a similar comparison for the Grade 3 item means for both numeracy (rows 'B' and 'D') and literacy (rows 'F' and 'H')

showed that the reverse was the case at the Grade 3 level (that is, the items appeared easier when taken by the NSW equating groups).

When interpreting the results provided in Table 6.3, it is necessary to remember that the data for the NSW item means were based on partial testing while the item means of SA data were based on a real test. Hence, it is necessary to make the following two assumptions.

1. The students sometimes did not try hard in the trial tests used for equating especially at the Grade 5 level. The 'did-not-try-hard' effect would seem to have worked more at Grade 5 where it could be assumed that students were old enough to tell the difference between the trial test and the real test and therefore to choose in which test to put more effort.
2. There was the possibility of a practice effect on a previous year test (for example, Grade 3 1997 Numeracy test), which occasionally made some tests appear much easier when taken by NSW equating students compared to when taken by the SA students. The practice effect would seem to have operated more at Grade 3 than at Grade 5.

In order to investigate the assumptions made above, further analyses were undertaken to examine how on the same scale the NSW equating groups performed in the trial test compared to the real test, and how the mean difficulty of the trial test compared with the mean difficulty of the real test taken by each NSW equating group.

These analyses involved concurrent equating the trial and the real tests taken by each of the NSW groups. Two important findings came out of the resulting analyses. First, there were far more students obtaining higher scores in the real tests compared to the trial test at both year levels regardless of the subject area. Second, it was evident that more often than not the equating students found the trial tests to be harder compared to the real test irrespective of the subject area. The latter observation can be made from Table 6.4 that compares the mean difficulties of the trial and the real tests obtained using the NSW equating groups. In 20 out of the 24 comparisons presented in Table 6.4, the trial test appeared harder than the real test.

Thus, the comparisons presented in Table 6.4 appear to support the assumption that the equating students sometimes 'did-not-try-hard' in the trial tests used for equating. Nevertheless, it should be noted that, although the assumption made here may have operated, it could not be distinguished from situations where there was marked difference in difficulty in the trial tests compared with the real test.

Some further exploratory analyses undertaken to investigate the effects of combining the NSW data with the SA data in concurrent equating the trial and the real tests, revealed that the SA data partially overruled the 'did-not-try-hard' error in the NSW data. However, concurrent equating in which SA data from the six occasions (1995 to 2000) was combined with NSW equating data gave results that were clearly inconsistent with expectations. The results indicated that the performance levels in SA for both numeracy and literacy in 1995 exceeded those of the other five occasions by at least 0.50 logits (that is, by about one year of learning). Clearly, something was wrong because it would be expected that the levels of performance in the State would generally not show such huge variations across the occasions.

Thus, it was evident that the lack of the 1994 NSW Grade 3 data coupled with the lack of SA data for 1994 to overrule the 'did-not-try-hard' error was distorting the equating results for 1995. Obviously, something needed to be done to avoid the 'did-not-try-hard' error in the link involving the 1994 NSW data (that is, the 1996-1994-1995 link) contaminating results for the other occasions in concurrent equating.

Table 6.4 Comparison of the mean difficulties of the trial and the real test using NSW equating groups

	Group Size (N)		**Test**		**N u m e r a c y**		**L i t e r a c y**	
Group	Grade 3	Grade 5			Grade 3	Grade 5	Grade 3	Grade 5
1994/1995	829	817	Trial	1994	0.08	0.25	0.18	0.09
			Real	1995	-0.08	-0.23	-0.19	-0.10
			Diff		0.16	0.48	0.37	0.19
1996/1994	NIL	976	Trial	1994	No Data	0.05	No Data	0.12
			Real	1996		-0.05		-0.12
			Diff			0.10		0.24
1996/1997	939	1,030	Trial	1996	0.17	0.02	0.11	0.09
			Real	1997	-0.17	-0.02	-0.11	-0.09
			Diff		0.34	0.04	0.22	0.18
1997/1998	1,154	1,204	Trial	1997	0.06	0.12	0.06	-0.04
			Real	1998	-0.06	-0.12	-0.06	0.04
			Diff		0.12	0.24	0.12	-0.08
1996/1999	1,000	974	Trial	1996	-0.23	0.10	0.11	-0.15
			Real	1999	0.20	-0.10	-0.17	0.15
			Diff		-0.43 Ψ	0.20	0.28	-0.30 Ψ
1999/2000	1088	1,126	Trial	1999	0.08	-0.02	0.04	-0.09
			Real	2000	-0.08	0.02	-0.04	0.09
			Diff		0.16	-0.04 Ψ	0.08	-0.18 Ψ

Notes:
Diff - the difference between the mean difficulty of the trial test and that of the real test
Ψ - The trial test appeared easier when compared with the real test

Consequently, in order to overcome the problem in the 1996-1994-1995 link as well as to minimize the 'did-not-try-hard' errors, the decision was made to follow three steps in equating across occasions:

(a) to use the concurrent technique to equate the combined NSW and SA data for 1996 to 2000 (This procedure would bring the SA data from 1996 to 2000 onto one common scale free from any possible contamination from the 1996-1994-1995 NSW equating data);

(b) to use NSW common students' differences in the 1996-1994-1995 Grade 5 data to equate the 1995 SA data to (a) above; and

(c) to make adjustments to the equating results obtained in (a) and (b) above so as to maintain the vertical distances between the Grades 3 and 5 tests obtained using the SA data alone as presented in Table 6.1.

It was considered logical to follow (c) above because with SA data the distances between the Grades 3 and 5 tests were estimated with huge numbers (21,000 to 25,000) of students and in real test situations (see Table 6.3). Thus, the distances calculated using SA data were considered more dependable compared to those calculated using the NSW data where about 1,500 to 4,500 students were used and half the data was from trial testing. Furthermore, it was considered logical to maintain

the magnitudes of the links obtained using items rather than those obtained using cases because case estimates were obviously less accurate than item estimates. For example, with NSW data, item estimates were calculated with about 1,000 students, while the case estimates were calculated with only 30 to 50 items.

Equating of the 1996 to 2000 tests

The cases in the combined 1996 to 2000 South Australia data plus the corresponding NSW equating data were 131,843 (123,328 cases from SA and 8,515 cases from NSW), which exceeded the limit of 100,000 cases that can be handled by the QUEST computer program. Consequently, for concurrent equating purposes, a decision was made to use half the cases from each testing occasion (1996 to 2000) from the South Australia data plus all the corresponding equating data from NSW.

Accordingly, SA data for each occasion was divided to form two samples. However, to avoid clustering error (since the students were ordered in the data files by schools), it was necessary to pick every other student from the data files to form one sample (called Set A) and the remainder of the students to form the other sample (called Set B). Table 6.5 gives the composition of the two sets after they were combined with the NSW equating data to form the concurrent data files. Set A contained 70,182 cases (61,667 and 8,515 cases from SA and NSW respectively), while Set B contained 70,176 cases (61,661 and 8,515 cases from SA and NSW respectively). Hence, the two samples had almost equal numbers of cases.

The calibration of the 1996 to 2000 tests was then done concurrently counting omitted and not-reached items as wrong, first using Set A then using Set B. All the items in Set A as well as in Set B numeracy data had INMS values within the 0.77 to 1.30 range showing they still had adequate fit to the Rasch model. However, Item 17 in the 1998 Grade 3 reading sub-test had INMS value of 1.33 (in both Sets A and B) and was therefore deleted bringing the total number of items deleted from the Literacy test to just three altogether.

The analyses carried out using Set A and those carried out using Set B gave more or less identical results and therefore, for reasons of parsimony, only Set A results are reported here.

Next, the thresholds of the 1996 to 2000 equated test items were anchored and used to calculate the scores of all the students at the Grade 3 (1996 to 2000) as well as at the Grade 5 (1996 to 2000) levels counting omitted items and not-reached items as wrong. The scores of students with perfect scores and those of students with zero scores were then estimated and the working rule of not calculating scores for students who did not respond to at least one item in the tests observed.

Equating of the 1995 test

Using only the NSW equating data and counting omitted and not-reached items as incorrect, the difference in mean difficulty of the (a) 1996 Grade 5 test and the 1994 Grade 5 test, and of the (b) 1994 Grade 5 test and the 1995 Grade 5 test were computed. The two differences from (a) and (b) were then summed up to calculate the relative difference in difficulty between the 1996 Grade 5 test and the 1995 Grade 5 test. This relative difference in difficulty between the 1996 Grade 5 test and the 1995 Grade 5 test was then used as an adjustment factor to bring the SA 1995 test onto the same scale with the 1996 to 2000 tests.

Table 6.6 presents the results of the comparison of the mean difficulties between the 1994, 1995 and 1996 Grade 5 Numeracy tests using NSW 1996/1994 and 1994/1996

equating groups. The symbol $\bar{X}$ followed by subscribed number is used here to represent the mean difficulty of the test for a particular year (for example, $\bar{X}_{94}$ stands for 'mean difficulty of the 1994 test').

Table 6.5 Composition of the equating sets

Group	Grade Level	Number of Cases		
		Set A	Set B	Set (A + B)
1996	Grade 3	5,548	5,547	11,095
	Grade 5	5,807	5,806	11,613
1997	Grade 3	6,219	6,218	12,437
	Grade 5	5,987	5,986	11,973
1998	Grade 3	6,397	6,397	12,794
	Grade 5	6,236	6,235	12,471
1999	Grade 3	6,275	6,275	12,550
	Grade 5	6,450	6,450	12,900
2000	Grade 3	6,339	6,338	12,677
	Grade 5	6,409	6,409	12,818
	Total (SA)	61,667	61,661	123,328
1996/1997	Grade 3	939	939	939
	Grade 5	1,030	1,030	1,030
1997/1998	Grade 3	1,154	1,154	1,154
	Grade 5	1,204	1,204	1,204
1996/1999	Grade 3	1,000	1,000	1,000
	Grade 5	974	974	974
1999/2000	Grade 3	1,088	1,088	1,088
	Grade 5	1,126	1,126	1,126
	Total (NSW)	8,515	8,515	8,515
	Total (SA + NSW)	70,182	70,176	131,843

From Table 6.6, it can be seen that using the NSW 1996/1994 equating group (N=976) the mean difficulty of the 1996 Grade 5 test is -0.05 logits and that of the 1994 test is 0.05 logits. Hence, the 1996 Grade 5 Numeracy test was on average easier by 0.10 logits compared to the 1994 Grade 5 Numeracy test. Similarly, using the NSW 1994/1995 equating group (N=817), the 1995 test was found on average to have been easier by 0.48 logits compared to the 1994 test.

Therefore, the 1995 test was on average easier than the 1996 test. Hence, in order to bring the 1995 numeracy scores to the same scale as the 1996 scores (and therefore, to the same scale as the 1996 to 2000 scores), the 1995 scores were to be dropped by a factor of 0.58 logits (that is, 0.10 + 0.48).

Similar calculations carried out for literacy (summary presented in Table 6.7) found that the 1995 literacy scores needed to be dropped by a factor of 0.46 to bring them onto the same scale as the 1996 to 2000 scores.

The adjustments for the 1995 numeracy and literacy tests were considered better estimates of the actual levels of the performance in SA compared to the estimates obtained using concurrent equating in which SA data from the six occasions (1995 to 2000) was combined with NSW equating data. It is highly likely that (a) the double link involved in connecting the 1995 tests to the 1996 tests, plus (b) the lack of the 1994 Grade 3 link, and (c) the lack of 1994 SA data were the causes of the obviously inflated 1995 results in the concurrent equating. The adjustment factors are considered to have provided the best alternative for circumventing these problems.

Table 6.6 Computation of the 1995 Numeracy test adjustment factor

	1996/1994 (N=976)	1994/1995 (N=817)
$\bar{X}_{94}$	0.05	0.25
$\bar{X}_{95}$		-0.23
$\bar{X}_{96}$	-0.05	
	$\bar{X}_{94} - \bar{X}_{96} =$ 0.10	$\bar{X}_{94} - \bar{X}_{95} =$ 0.48

Table 6.7 Computation of the 1995 Literacy test adjustment factor

	1996/1994 (N=976)	1994/1995 (N=817)
$\bar{X}_{94}$	0.12	0.09
$\bar{X}_{95}$		-0.10
$\bar{X}_{96}$	-0.15	
	$\bar{X}_{94} - \bar{X}_{96} =$ 0.27	$\bar{X}_{94} - \bar{X}_{95} =$ 0.19

Adjustment of the equating results

Tables 6.8 and 6.9 present the equating results for numeracy and literacy respectively. In Panel 1 of the two tables, the vertical equating results[15] for the six occasions obtained using the SA data are presented and the differences of the case means as well as item means between the grade levels are indicated (Tables 6.8 and 6.9, differences marked as 'A' and 'B'). Panel 2(a) of both tables present the concurrent equating results for 1996-2000 using half the SA data combined with all the corresponding NSW equating data (N=70,182). Panel 2(b) of the tables presents the final overall equating results for the six occasions after making adjustments to:

(i) bring the 1995 test to the same scale with the other five (1996-2000) tests using the adjustment factors computed earlier in Tables 6.6 and 6.7 for numeracy and literacy respectively, and

(ii) maintain the differences between the Grades 3 and 5 tests, and consequently the case means observed in vertical equating using SA data only.

Equating checks

In order to establish consistency of equating results obtained using NSW data combined with SA data, some checks were developed. As a result, Check 1 in the Tables 6.8 and 6.9 compares the differences between the Grades 3 and 5 case means

[15] These are the same results presented above (Table 6.2) but are reproduced here to make comparison easier.

obtained in vertical equating using SA data only (difference 'A') and the differences between the two means when using half SA data combined with the NSW equating data (difference 'C'). In the same way, Check 2 compares the differences between the item means obtained in the vertical equating (difference 'B') and those obtained in the overall equating (difference 'D').

Table 6.8 The final equating results for numeracy

1. Within occasion equating using all SA data

	1995	1996	1997	1998	1999	2000
Case Mean						
Grade 5	1.79	1.19	1.20	1.40	1.30	1.28
Grade 3	0.88	0.26	0.10	0.28	0.14	0.19
Difference (A)	0.91	0.93	1.11	1.12	1.16	1.09
Item Mean						
Grade 5	0.26	0.48	0.45	0.39	0.33	0.31
Grade 3	-0.45	-0.96	-0.70	-0.65	-0.62	-0.58
Difference (B)	0.71	1.44	1.15	1.04	0.95	0.89

2. Overall equating using NSW equating data plus half SA data

a) Before adjustment

Case Mean						
Grade 5		1.24	1.31	1.34	1.36	1.24
Grade 3		0.30	0.08	0.19	0.18	0.15
Difference (C)		0.94	1.23	1.15	1.18	1.09
Check 1 {=A-C}		-0.01	-0.13	-0.03	-0.02	-0.02
Item Mean						
Grade 5		0.54	0.56	0.33	0.39	0.24
Grade 3		-0.94	-0.71	-0.75	-0.54	-0.67
Difference (D)		1.48	1.27	1.08	0.93	0.91
Check 2 {=B-D}		-0.04	-0.12	-0.04	0.02	-0.02

b) After adjustment

Case Mean						
Grade 5 (Final)	***1.21***	1.24	1.31	1.34	1.36	1.24
Grade 3 (Final)	***0.30***	***0.31***	***0.21***	***0.22***	***0.20***	***0.15***
Difference (E)	0.91	0.93	1.10	1.12	1.16	1.09
Check 3 {=A-E}	0.00	0.00	0.00	0.00	0.00	0.00
Growth			1.01	1.03	1.15	1.02
Item Mean						
Grade 5 (Final)	0.26	0.54	0.56	0.33	0.39	0.24
Grade 3 (Final)	0.45	***-0.90***	***-0.59***	***-0.71***	***-0.56***	***-0.65***
Difference (F)	0.71	1.44	1.15	1.04	0.95	0.89
Check 4 {=B-F}	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

Table 6.9 The final equating results for literacy

1. Within occasion equating using all SA data

	1995	1996	1997	1998	1999	2000
Case Mean						
Grade 5	1.82	1.54	1.74	1.57	1.15	1.10
Grade 3	0.84	0.36	0.48	0.29	0.22	-0.01
Difference (A)	0.98	1.19	1.26	1.28	0.93	1.11
Item Mean						
Grade 5	0.23	0.23	0.29	0.38	0.21	0.21
Grade 3	-0.62	-0.52	-0.55	-0.62	-0.28	-0.48
Difference (B)	0.85	0.75	0.84	1.00	0.49	0.69

2. Overall equating using NSW equating data plus half SA data

a) Before adjustment

Case Mean						
Grade 5		1.48	1.57	1.50	1.31	1.29
Grade 3		0.34	0.30	0.22	0.34	0.19
Difference (C)		1.14	1.27	1.28	0.97	1.10
Check 1 {=A-C}		**0.04**	**-0.01**	**0.00**	**-0.04**	**0.01**
Item Mean						
Grade 5		0.19	0.19	0.31	0.38	0.40
Grade 3		-0.54	-0.73	-0.73	-0.15	-0.29
Difference (D)		0.73	0.92	1.04	0.53	0.69
Check 2 {=B-D}		**0.02**	**-0.08**	**-0.04**	**-0.04**	**0.00**

b) After adjustment

Case Mean						
Grade 5 (Final)	***1.36***	1.48	1.57	1.50	1.31	1.29
Grade 3 (Final)	***0.38***	***0.30***	***0.31***	0.22	***0.38***	0.18
Difference (E)	0.98	1.18	1.26	1.28	0.93	1.11
Check 3 {=A-E}	0.00	0.00	0.00	0.00	0.00	0.00
Growth			1.19	1.20	1.00	1.07
Item Mean						
Grade 5 (Final)	0.23	0.19	0.19	0.31	0.38	0.40
Grade 3 (Final)	-0.62	***-0.56***	***-0.65***	***-0.69***	***-0.11***	-0.29
Difference (F)	0.85	0.75	0.84	1.00	0.49	0.69
Check 4 {=B-F}	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

Judging from Check 1 and 2, there were obviously very small deviations (less than 0.10 for the majority) of the item and case means after the overall equating compared to what was obtained in vertical equating. The deviations recorded here were particularly small compared to those observed earlier when vertical equating results using SA data were compared with those obtained using NSW equating data (see Table 6.3).

Clearly, in concurrent equating where the NSW data are combined with the SA data, the SA data seem to overrule partially the predicted 'did-not-try-hard' error in the NSW data. For example, the largest deviation while dealing with the separate NSW and SA data was -0.30 (Table 6.3) for the 1997 Numeracy test and this deviation (though still the largest) had dropped by more than half to just -0.13 (Table 6.8, Check 1) when dealing with the combined data.

Although a majority of the deviations recorded as Checks 1 and 2 (Tables 6.8 and 6.9) were small (that is, less than a term's schoolwork), they nevertheless were considered as errors due to the effects of students sometimes 'not trying hard' in the trial tests used for equating. Consequently, the item and case means were adjusted to remove the deviations observed in Checks 1 and 2. So, Checks 3 and 4 are akin to Checks 1 and 2 respectively but after adjustments have been made to the case and item means to maintain the distances between the Grades 3 and 5 observed in vertical equating using SA data only. The values of the adjusted case and item means are given in *Italics* in the two tables.

Effects of the double links in concurrent equating

As illustrated above in Figure 6.1, double links were involved in connecting the 1995, 1998 and 2000 tests to the 1996 tests. However, as previously outlined, the 1995 tests were dropped from concurrent equating leaving the 1998 and 2000 as the only tests linked indirectly to the 1996 tests in the concurrent file.

From Tables 6.8 and 6.9 (in the rows giving the final case mean), it can be observed that a drop in performance in SA appeared to have occurred whenever there was a double link (except for the 1998 Numeracy test). For instance, there were drops in numeracy achievement going from 1999 to 2000 at both year levels. Similarly, there were drops in literacy achievement going from 1997 to 1998, and again going from 1999 to 2000 at both year levels. However, there were no such patterns of achievement drops observed in most of the tests that were linked directly to the 1996 test. Although it was difficult to tell whether the double links were the problems or whether the levels of achievement had actually dropped in those situations, the observed patterns raise concerns regarding the appropriateness of using double links to equate the tests.

If the assumption that the students sometimes do not try hard in the trial tests used for equating, then larger equating errors would be expected to occur in situations involving double links than in situations involving direct links because two sets of trial tests are involved in the former situation. There are clear possibilities that the double links used in the equating design might have distorted the equating results to some extent but in a way that is not readily understood.

Levels of achievement across occasions

Regardless of the fact that the equating might have been distorted by the problems outlined above, the final case and item means presented in Tables 6.8 and 9 must be considered as the best plausible estimates of the actual levels of achievement in numeracy and literacy in SA. This is because the procedure used was logical, and it gave achievement levels that did not show huge (and therefore impossible) variations year after year.

Figure 6.2 and 6.3 show the achievement levels for numeracy and literacy respectively, across the six testing occasions at the Grades 3 and 5 levels in South Australia.

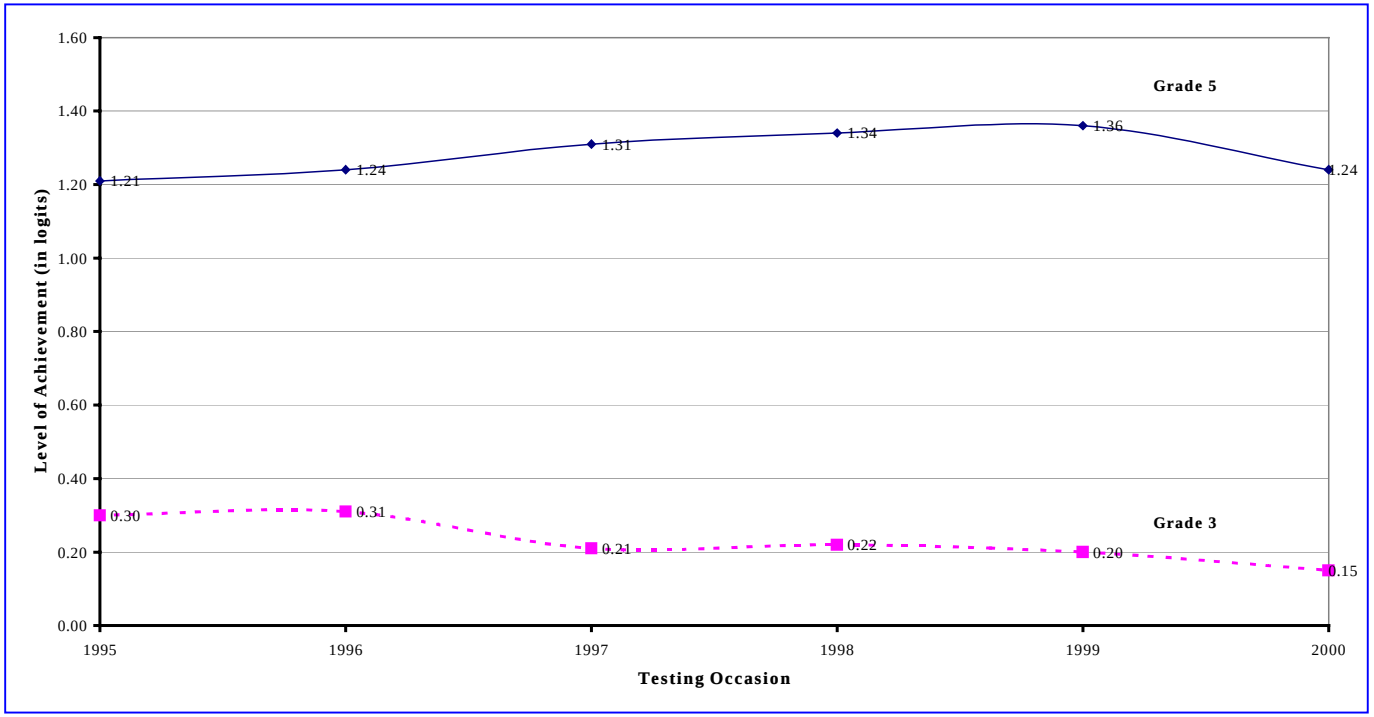


Figure 6.2 1995 to 2000 Grades 3 and 5 levels of numeracy achievement

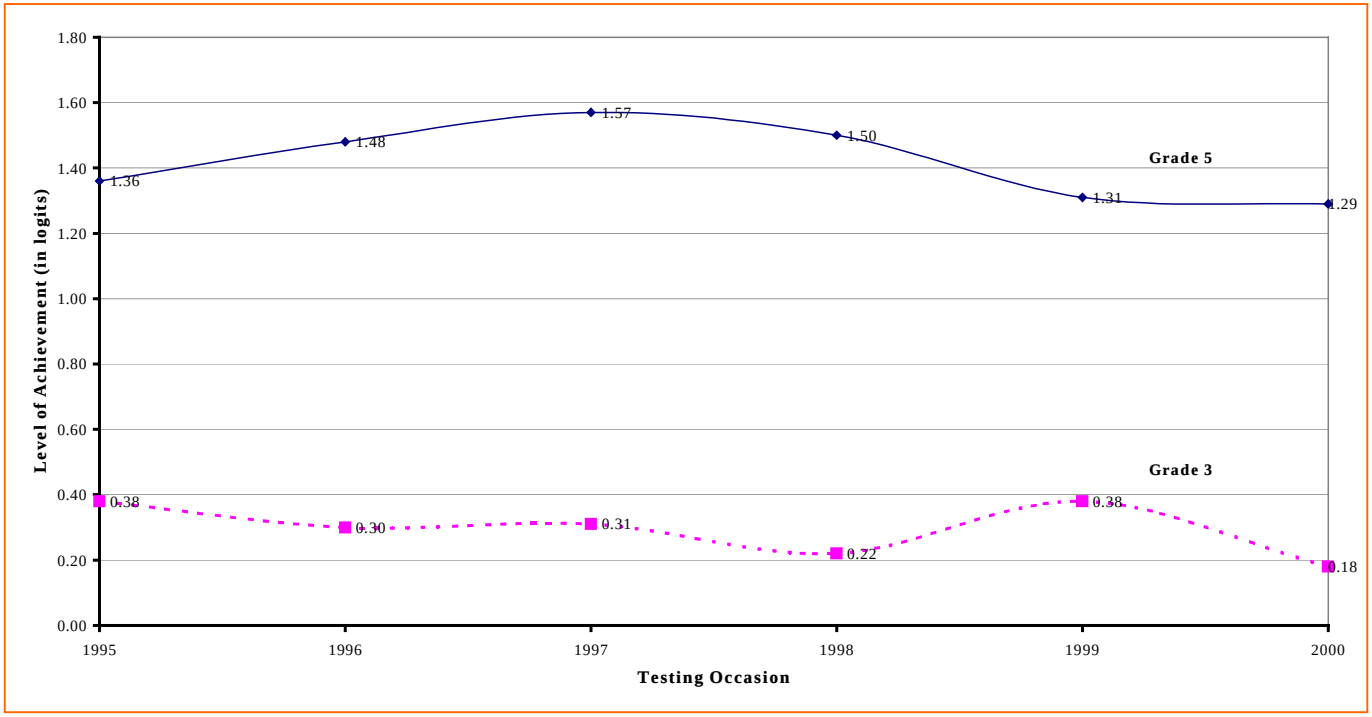


Figure 6.3 1995 to 2000 Grades 3 and 5 levels of literacy achievement

---

Note: In Figures 6.2 and 6.3, a change in achievement of ±0.13 between occasions should be considered substantial because it represents about one term of school learning (Hungi, 1997).

From Table 6.8 and Figure 6.2 it can be observed that at the State level, the achievement of the Grade 5 students in numeracy increased slightly but steadily each year from 1995 (1.21) to 1999 (1.36), and then a slight drop occurred in 2000 (1.24). For the Grade 3 students, the achievement level in numeracy in 1995 and 1996 was around 0.30 logits. However, this achievement level of the Grade 3 students decreased by a substantial amount to 0.21 in 1997. The performance then seemed to stabilize at around 0.20 for the next years but a slight (0.05) drop occurred in 2000 resulting in a State average of 0.15 for that year. Perhaps the substantial drop in the average Grade 3 numeracy achievement after 1995 and 1996 could be a consequence of the large increase in the number of students taking part in the BSTP on the later testing occasions.

On the other hand, from Table 6.9 and Figure 6.3 it can be observed that at the State level, the achievement of the Grade 5 students in literacy increased substantially for the first three years of testing then dropped for the next three occasions. That is, from 1.36 in 1995 to 1.48 and 1.57 in 1996 and 1997 respectively, to 1.50, 1.31 and 1.29 for 1998, 1999 and 2000, respectively. However, the achievement levels of the Grade 3 students in literacy seem not to have followed any identifiable pattern. Nevertheless, on four out of the six testing occasions, the literacy achievement averages at the State level for the Grade 3 students were recorded as 0.35±0.05 and therefore it could be argued that the State average dropped by a substantial amount for 1998 (0.22) and 2000 (0.18).

Growth in achievement between Grades 3 and 5

The growths recorded in Tables 6.8 and 6.9 represent the average increase in numeracy and literacy achievement respectively between Grades 3 and 5 in the same set of students in South Australia. For instance, to obtain the first growth value recorded under numeracy (1.01), the mean performance of the students at Grade 3 in 1995 (0.30) was subtracted from their mean performance at Grade 5 in 1997 (1.31). It should be noted that the growth values calculated here are just general estimates because they do not take into account absenteeism from the tests or transience in and out of the State. However, it was interesting to note that the values compared well with those computed using vertical equating within the same occasion. It appears that the growth in achievement in the basic skills of numeracy and literacy between Grade 3 and 5 on average is somewhere around 0.50 logits per year. Consequently, the differences reported and discussed in the previous paragraphs should be interpreted with the knowledge that 0.50 logits represents the advancement in both numeracy and literacy during a year of schooling at the mid-primary school stage.

Conclusions and Recommendations

In summary, the analyses reported in this chapter have brought to light several technical points that must be taken into account when equating the South Australia BST data across occasions.

1. It could be assumed that students sometimes do not ‘try their hardest’ in the trial tests used for equating especially at the Grade 5 level.
2. In order to minimize distortion in equating results caused by the error in (1) above in the NSW data, the following steps are necessary.
    (a) The concurrent equating technique where NSW data and SA data are combined must be applied because the SA data partially overrules the error in the NSW data.

(b) The relative vertical distances between the Grades 3 and 5 data in SA must be maintained in the final across occasions equating results obtained using SA data combined with NSW equating data. This is because the distances obtained using SA data are more dependable than those obtained using NSW data since the latter are based on trial testing; while the former are based on the real testing program.

3. Although (1) might operate, it can not be distinguished from situations where there is marked difficulty in the tests.
4. There is the possibility of a practice effect on a previous year test (for example, Grade 3 1997/1996 test). The practice effect seems to work at Grade 3.
5. There are some doubts concerning the appropriateness of using double links in connecting the 1995, 1998 and 2000 tests to the 1996 test. A preferred procedure would be to link all the tests directly to one common test that must be kept secure to eliminate any errors that may results from double links.

By and large, the analyses presented in this chapter have raised issues regarding the appropriateness of the procedures of using common students for across occasions equating purposes. Given that the error in (1) above is almost inevitable, a preferred procedure would be for the test developers to include some common items in the tests across occasions so as to allow more accurate linking over time. In addition, the test developers would need to keep the tests secure to avoid students on future occasions obtaining access to the common items.

The analyses reported in this chapter also bring to light information regarding the scaling characteristics of the items in the BSTP. Overwhelmingly, the items had adequate fit to the Rasch model and the item means between Grades 3 and 5 compare well year after year. Clearly, the test developers did excellent work in the development of the items and in allocation of the items to either the Grade 3 or the Grade 5 tests.

This study has also brought to light information regarding levels of achievement as well as growth in achievement in numeracy and literacy in South Australia government primary schools Grades 3 and 5 students. With only a few exceptions, the achievement in both numeracy and literacy at the Grade 5 grade has continued to increase since the inception of the program six years ago in 1995. However, the achievement in numeracy and literacy at the Grade 3 level has remained fairly constant. Finally, the growth in achievement between Grades 3 and 5 for both subjects has remained approximately 0.50 logits per year. However, this growth has continued to increase slightly year after year especially for numeracy.

7

Achievement Factors: Two-level Models

This chapter reports the two-level HLM analyses carried out to examine factors influencing achievement in numeracy and literacy among Grades 3 and 5 primary school students in South Australia. The three two-level models (namely, Model-X, Model-Y and Model-Z) that are proposed in Chapter 5 for teasing out factors influencing students' achievement in the BST are examined in this chapter. For each type of model proposed, two-level HLM analyses are undertaken to:

(a) examine the amounts of variances available at the student-level and at the school-level;

(b) examine the amount of variance explained in the model by prior achievement alone;

(c) examine the amounts of variances explained by the predictors in the final model at the student-level and at the school-level;

(d) examine the goodness of fit of the model based on deviance statistic and the chi-square test, and

(e) identify the student-level and school-level factors together with the interaction effects involved in students' achievement in the basic skills of Numeracy and Literacy across Grade 3 and 5 primary school grade levels in South Australia.

For comparison purposes, the results of the analyses of the proposed models are presented together. The HLM5/2L computer program developed by Raudenbush, Bryk and Congdon (2000) is used to carry out all the multilevel analyses reported in this chapter.

Descriptions of the two-level HLM models

The three two-level hierarchical linear models (Model-X, Model-Y and Model-Z) proposed for teasing out factors influencing students' achievement in this study are

introduced in Chapter 5 (Figures 5.2, 5.3 and 5.4 respectively). The data sets and variables examined for inclusion in each of the three models are also described in that chapter and therefore it is unnecessary to repeat those details here. Nevertheless, it should be borne in mind that for each of the three types of models proposed, two separate models are specified, one for numeracy and the other for literacy.

Specifications of the two-level null models

Raudenbush et al. (2000) have recommended the running of a variance decomposition (fully unconditional or null) model as a starting point in hierarchical data analysis, as it provides useful information about the outcome variability at the different levels of the hierarchy. The estimates of variance to be explained are obtained from the null model, and may then be used to calculate the amount of variance explained by the final model. The null model is the simplest model because no predictor variables are specified at any level. However, for Model-X, the variable YEARLEVL (Year of Study or Grade Level) is included in the simplest model in order to differentiate between Grade 3 and Grade 5 students.

For Models Y and Z, and following the procedure as well as the symbols given by Raudenbush and Bryk (2002; pp.69-70), the simplest two-level models to represent how variation in the outcome variables is allocated across the two different levels (student and school), can be stated as follows.

Level-1 model

At the student-level, the student achievement is modelled as a function of a school mean plus a random error:

$$\mathbf{Y}_{ij} = \boldsymbol{\beta}_{0j} + \mathbf{r}_{ij} \qquad \text{Equation 7.1}$$

where:

$\mathbf{Y}_{ij}$ is the achievement (Rasch score) of student *i* at Grade 5 in school *j*;

$\boldsymbol{\beta}_{0j}$ is the mean achievement of school *j*; and

$\mathbf{r}_{ij}$ is a random error or 'student effect', that is, the deviation of the student mean from the school mean.

The indices *i* and *j* denote students and schools where there are

$i = 1, 2, \ldots, n_j$ students within school j; and

$j = 1, 2, \ldots, J$ schools.

A simplified form of Equation 7.1 is presented in the output file generated by the HLM5/2L computer program, where **Y**, **B0** and **R** are used to represent the components Y_{ij}, β_{0j} and r_{ij} in Equation 7.1, respectively. Hence, the Level-1 null model equation in the output file becomes:

$$\mathbf{Y} = \mathbf{B0} + \mathbf{R} \qquad \text{Equation 7.2}$$

Level-2 model

At the school-level, each school mean, $\boldsymbol{\beta}_{0j}$, is viewed as an outcome varying randomly around some grand mean:

$$\boldsymbol{\beta}_{0j} = \boldsymbol{\gamma}_{00} + \mathbf{u}_{0j} \qquad \text{Equation 7.3}$$

where:

$\boldsymbol{\beta}_{0j}$ is the mean achievement of school *j*; and

γ_{00} is the grand mean,

$\mathbf{u}_{0j}$ is a random 'school effect', that is, the deviation of the school mean from the grand mean. Within each of the schools, the variability among students is assumed to be the same.

A simplified form of Equation 7.3 that is presented in the output file generated by the HLM5/2L computer program is:

B0 = G00 + U0 Equation 7.4

where:

B0, **G00** and **U0** are used to represent the components β_{0j}, γ_{00} and $\mathbf{u}_{0j}$ in Equation 7.3, respectively.

For Model-X, using the symbols introduced above, the simplest two-level models with the variable YEARLEVL as the only predictor, can be stated as follows.

Level-1 Model

Y = B0 + B1*(YEARLEVL) + R Equation 7.5

where:

Y is the achievement (Rasch score) of the student at Grades 3 and/or 5;

B0 is the mean achievement of the school, that is, the intercept;

B1 is the mean growth rate in achievement of the school, that is, the regression slope associated with YEARLEVL; and

R is a random error.

Hence, for the grade-level-only model, the Level-1 model involves the estimation of two coefficients for each school: the intercept and the YEARLEVL slope.

In order to facilitate the interpretation of the fixed effects, in the HLM analyses, the variable YEARLEVL is entered into the equations uncentred.

Level-2 Model

The two Level-1 coefficients described above become outcome variables at Level-2 (school-level).

B0 = G00 + U0 Equation 7.6

B1 = G10 + U1 Equation 7.7

where:

G00 is the grand mean,

G10 is the mean growth rate in achievement of the schools,

U0 is a random 'school effect', that is, the deviation of the school mean from the grand mean.

U1 is a random 'year-of-study effect', that is, the deviation of the school mean growth rate from the grand mean growth rate.

Hence, at Level-2 of the model, each Level-1 coefficient is modelled as randomly varying among the schools.

Variance partitioning

Table 7.1 displays estimates of the variance involved in the two-level models for numeracy and literacy. The percentages of variance available at each of the two levels of hierarchy are calculated from the variance components by employing the formulae presented in Chapter 4.

Table 7.1 Variance partitioning based on the two-level models

	Term	**Model-X**[a]		**Model-Y**[b]		**Model-Z**[c]	
		Variance Component	(%) Var. Available	Variance Component	(%) Var. Available	Variance Component	(%) Var. Available
Numeracy							
Student	σ_0^2	1.17	(75.1)	1.00	(81.7)	0.99	(82.1)
School	$\tau_{\pi 0}$	0.39	(24.9)	0.22	(18.3)	0.21	(17.9)
Total	$\sigma_0^2 + \tau_0$	**1.56**		**1.22**		**1.20**	
Literacy							
Student	σ_0^2	1.21	(79.0)	1.00	(81.8)	0.97	(81.7)
School	$\tau_{\pi 0}$	0.32	(21.0)	0.22	(18.2)	0.22	(18.3)
Total	$\sigma_0^2 + \tau_0$	**1.53**		**1.22**		**1.19**	

Note: a - For Model-X, the simplest model has the variable YEARLEVL as the only predictor.
b - Transience model
c - Non-transience model

In interpreting the results displayed in Table 7.1 for Model-X, it should be remembered that the null models specified for numeracy and literacy have the variable YEARLEVL included to differentiate between Grade 3 and Grade 5 students. It should also be borne in mind that the null models of the other two types of models (that is, Models Y and Z) have no predictors included. By definition, the simplest models specified for Model-X are not true null models. Therefore, the results of variance partitioning based on Models Y and Z provide better pictures of the situations for student scores than the results based on Model-X. Nevertheless, it is interesting to note that the results displayed in Table 7.1 for Model-X follow closely the results for Models Y and Z, which indicate that the variations of the students' scores for the two grades are nearly the same.

Hence, the results in Table 7.1 show that, based on the transience model (Model-Y), 81.7 and 18.3 per cent of the variation of Grade 5 pupils' numeracy scores are at the student and school levels respectively, and based on the non-transience model (Model-Z) the percentages are 82.1 and 17.9 respectively. The corresponding percentages for students' literacy scores based on the transience model are 81.8 and 18.2 for student and school-level respectively, and the corresponding percentages based on the non-transience model are 81.7 and 18.3 respectively. These variations of students' scores at the two levels of hierarchy are the maximum amounts of variance available at those levels that can be explained in subsequent analyses.

For both outcome measures, the variation of students' scores at the student-level and at the school-level based on the transience model are basically the same as the variation at those levels based on the non-transience model. These results indicate that, at the Grade 5 level in South Australia, the inclusion of the transience students in these two-level analyses does not noticeably alter the variations of students' scores in numeracy and literacy. In addition, the results of variance partitioning for numeracy follow

closely those for literacy, which indicate that, at Grade 5 primary school level in South Australia, the variations of the students' scores for numeracy and literacy are nearly the same.

Thus, the results for the transience and non-transience models displayed in Table 7.1 indicate that in South Australia, the variation between students within schools in terms of their achievement in numeracy and literacy at Grade 5 is roughly around four times greater when compared with the variation in performance between schools. That is, there is huge variability within the schools when compared to the variability between the schools.

Finally, it should be noted that the results of variance partitioning reported above are consistent with the results reported by Afrassa and Keeves (1999) in their two-level HLM analyses of the 1995 to 1997 South Australian BSTP data sets. Afrassa and Keeves analyzed data for each grade level (Grade 3 and Grade 5) separately, and from each of the three testing occasions (1995 to 1997) separately. They found that for both grade levels and for both outcome measures (numeracy and literacy) the variations of students' scores at student-level were roughly around four times when compared to variation in performance between schools.

Effects of grade level

In this sub-section, the results of fixed effects from the analyses of the so-called 'grade-level-only' model (that is, the simplest form of Model-X) are reported. Importantly, the main aim of the results presented in this sub-section is to examine the mean growth in achievement between the two grades without including all other factors. Consequently, the variable YEARLEVL is added as the only predictor in Model-X and no other predictors are specified at Levels 1 and 2. The general equations involved in the grade-level-only model are presented above (Equations 7.5 to 7.7).

The final estimation of fixed effects for the grade-level-only models for numeracy and literacy are presented in Table 7.2. In interpreting the results, it should be remembered that variable YEARLEVL is entered into the equations uncentred.

The results in Table 7.2 indicate that variable YEARLEVL has significant ($p<0.05$) influences on achievement in both numeracy and literacy. The positive coefficients and positive t-values for YEARLEVL indicate that Grade 5 students are likely to perform better (in either numeracy or literacy) than their Grade 3 counterparts. For numeracy, the table shows a significant average growth rate of 0.52 logits per year, (with a t-value of 105.09), while for literacy, the table indicates a significant average growth rate of 0.56 logits per year, (with a t-value of 121.04).

In order to obtain the average growth between the two grades, the average growth rates recorded in Table 7.2 should be doubled, since the predictor YEARLEVL was coded (Grade 3 = 0, Grade 5 = 2). Hence, on average growth in numeracy and literacy achievements between Grades 3 and 5 in South Australia primary schools is estimated to be 1.05 logits and 1.12 logits, respectively. However, it should be emphasized that no student and no school characteristics have been considered in the calculation of the growth rates obtained here.

From the final estimates of the fixed effects provided in Table 7.2, it can also be noted that the estimated intercept (grand mean) for numeracy is 0.17 and that of literacy is 0.21, with significant t-ratios of 13.38 and 17.69, respectively. The numeracy and the literacy scores used here are Rasch scaled and that zero is the mean difficulty level of the items in the 1996 tests (see Chapter 6). Therefore, the t-tests here indicate that the

mean scores for Grade 3 differ from the mean difficulty levels of the items in the 1996 tests. In general terms, the difficulty level of the items included in the tests were slightly below the performance levels on average of the students at the Grade 3 level.

Table 7.2 Final estimation of fixed effects for the grade-level-only models

Model-X		Coefficient	Std. Error	T-ratio	P-value
Numeracy					
For **INTRCPT1**, B0					
INTRCPT2, G00		0.17	0.01	13.38	0.00
For YEARLEVL slope, B1					
INTRCPT2, G10		0.52	0.00	105.09	0.00
Literacy					
For **INTRCPT1**, B0					
INTRCPT2, G00		0.21	0.01	17.69	0.00
For **YEARLEVL** slope, B1					
INTRCPT2, G10		0.56	0.00	121.04	0.00

Effects of prior achievement

This sub-section reports on the analyses carried out to examine the effects of Prior Achievement (Grade 3 score) on the achievement of the students at Grade 5 in numeracy and literacy, without including any other factors. In order to achieve this aim, the variable Y3NSCORE (for numeracy) and Y3LSCORE (for literacy) are added in Models Y and Z as the only predictors in Equation 7.2 presented above. No other predictors are specified at Levels 1 and 2.

For example, by adding Y3NSCORE as a predictor in either Model-Y or Model-Z, the two-level HLM prior-achievement-only model for numeracy is specified in equation format as follows:

Level-1 model

$$\mathbf{Y = B0 + B1*(Y3NSCORE) + R} \qquad \text{Equation 7.8}$$

Level-2 model

$$\mathbf{B0 = G00 + U0} \qquad \text{Equation 7.9}$$

$$\mathbf{B1 = G10 + U1} \qquad \text{Equation 7.10}$$

The final estimation of fixed effects obtained from Models Y and Z for the prior-achievement-only models for numeracy and literacy are presented in Table 7.3.

The results in Table 7.3 indicate that Prior Achievement has a significant influence on achievement in numeracy as well as literacy, and in both Model-Y and Model-Z. The positive coefficients and t-ratio values for Y3NSCORE (or Y3LSCORE) indicate that Grade 5 students who were high achievers at Grade 3 are likely to achieve better (in either numeracy or literacy) than their Grade 5 counterparts who were low achievers at Grade 3.

For numeracy, Table 7.3 shows a significant positive effect of Prior Achievement on the final score of 0.57 and 0.59 logits for Models Y and Z respectively. For literacy, the table indicates that the effect of Prior Achievement on the final score is 0.63 and 0.64 logits for Models Y and Z respectively. It should be noted that the coefficient for

Prior Achievement in Models Y and Z are almost equal for the same subject. Hence, it appears that the effect of Prior Achievement is almost the same whether only those students who remain in the same school are considered (Model-Z) or all matched students are considered (Model-Y).

Table 7.3 Final estimation of fixed effects for the prior-achievement-only models

	Model - Y				Model - Z			
	Coeff.	Std. Error	T-ratio	P-value	Coeff.	Std. Error	T-ratio	P-value
Numeracy								
For **INTRCPT1**, B0								
INTRCPT2, G00	1.37	0.01	157.97	0.00	1.42	0.01	159.67	0.00
For **Y3NSCORE** slope, B1								
INTRCPT2, G10	0.57	0.01	124.95	0.00	0.59	0.01	119.61	0.00
Literacy								
For **INTRCPT1**, B0								
INTRCPT2, G00	1.48	0.01	181.29	0.00	1.53	0.01	184.79	0.00
For **Y3LSCORE** slope, B1								
INTRCPT2, G10	0.63	0.00	159.56	0.00	0.64	0.00	152.20	0.00

The results of the final estimations of the variance components for the prior-achievement-only models and the results of the variance components obtained from the null models (provided in Table 7.1) are presented together in Table 7.4 (rows marked 'a' and 'b') for ease of comparison.

From the information in Table 7.4 rows 'a' and 'b', the percentages of total variance explained by Prior Achievement at the two levels is calculated and the results of the calculations are presented in rows 'e' of the table. A description of the procedure undertaken to calculate the values presented in Table 7.4 for percentages of variances explained by Prior Achievement is presented in Chapter 4.

The results in Table 7.4 show that, for numeracy, the percentages of total variances explained by Prior Achievement at the two levels are 46.2 and 46.9 for the transience (Model-Y*)* and non-transience (Model-Z) models respectively. The corresponding percentages for literacy are 55.7 and 56.6 for the transience and non-transience models respectively.

The percentages of total variance that are explained by Prior Achievement in the transience model and the non-transience model are almost equal for the same subject. Hence, it seems that that Prior Achievement accounts for almost the same amount of variance whether only those students who remain in the same school are considered (Model-Z) or all matched students are considered (Model-Y). In addition, the amounts of variance explained by Prior Achievement in both Models Y and Z are large at the school-level (roughly around 60 per cent) indicating that most variation between the schools is accounted for by Prior Achievement alone when change is assessed over a two-year period.

Two-level unconditional models

The next step in the analyses is to model achievement in numeracy and literacy as the outcome variables predicted by the Grade Level (in Model-X) or Prior Achievement and Transience (in Model-Y) or Prior Achievement (in Model-Z) together with other

student background variables. No predictors are specified at Level-2, and therefore, Raudenbush et al. (2000) have referred to this type of model as 'unconditional' at Level-2.

Table 7.4 Variance explained by Prior Achievement

	Model - Y			Model - Z		
	Level-1	Level-2	Total	Level-1	Level-2	Total
Numeracy						
a) Var. Comp. Null Model	1.00	0.22	1.22	0.99	0.21	1.20
b) Var. Comp. Prior-only Model	0.57	0.09		0.55	0.09	
c) Var. Available	81.7%	18.3%		82.1%	17.9%	
d) Var. Explained by Prior Ach.	43.1%	59.8%		44.5%	58.1%	
e) Total Var. Explained by Prior Ach.	35.2%	11.0%	46.2%	36.5%	10.4%	46.9%
Literacy						
a) Var. Comp. Null Model	1.00	0.22	1.22	0.97	0.22	1.19
b) Var. Comp. Prior-only Model	0.46	0.08		0.44	0.08	
c) Var. Available	81.8%	18.2%		81.7%	18.3%	
d) Var. Explained by Prior Ach.	54.2%	62.8%		55.2%	62.9%	
e) Total Var. Explained by Prior Ach.	44.3%	11.4%	55.7%	45.1%	11.5%	56.6%

It was mentioned in Chapter 5 that at the student-level, six variables (AGE, SEX, ATSI, HOME, NESB and INOZ) are available that can be examined for their influence on achievement in numeracy and literacy in all the three proposed models. However, only a maximum of five of the six variables can be entered into the equations at any one time. This is because the variables HOME and NESB are alternative versions of the same measure and therefore should not be entered into the equation simultaneously (see Chapter 3).

A step-up approach is followed to examine which of the student-level variables have a significant influence on achievement in numeracy and literacy in each of the proposed models. Bryk and Raudenbush (1992) have recommended the step-up approach of inclusion of variables into the model to the alternative step-down approach where all the possible predictors are included in the model and then the non-significant variables are progressively eliminated from the model.

The final Level-1 unconditional models for numeracy and literacy are presented in Equations 7.11 to 7.15 for Models X, Y and Z. For Model-X, the unconditional models for numeracy and literacy are similar and are therefore reported together below.

Model-X

For both numeracy and literacy

$$Y = B0 + B1*(YEARLEVL) + B2*(AGE) + B3*(ATSI) + B4*(HOME) + R \quad \text{Equation 7.11}$$

Model-Y

For numeracy

$$Y = B0 + B1*(SEX) + B2*(TRANS) + B3*(AGE) + B4*(ATSI) + B5*(INOZ) + B6*(Y3NSCORE) + R \quad \text{Equation 7.12}$$

For literacy

$$Y = B0 + B1*(SEX) + B2*(TRANS) + B3*(AGE) + B4*(ATSI) + B5*(HOME) + B6*(INOZ) + B7*(Y3LSCORE) + R \quad \text{Equation 7.13}$$

Model-Z

For numeracy

$$Y = B0 + B1*(SEX) + B2*(AGE) + B3*(ATSI) + B4*(INOZ) + B5*(Y3NSCORE) + R \quad \text{Equation 7.14}$$

For literacy

$$Y = B0 + B1*(SEX) + B2*(AGE) + B3*(ATSI) + B4*(HOME) + B5*(INOZ) + B6*(Y3LSCORE) + R \quad \text{Equation 7.15}$$

For Model-X, Equation 7.11 indicates that at the student-level, the variables that have a significant influence on the outcome variables are four, namely YEARLEVL, AGE, ATSI and HOME. The Gender of the Student (SEX) and the Migrant Status in Australia (INOZ) variables have no significant influence on the outcome variables in Model-X.

For Models Y and Z, Equations 7.12 to 7.15 indicate that the five student-level variables have a significant influence on both numeracy and literacy. These five variables are namely SEX, AGE, ATSI, INOZ and Prior Achievement, that is, either Y3NSCORE or Y3LSCORE for numeracy and literacy respectively. In addition, Equations 7.12 to 7.15 indicate that the variable HOME has a significant influence on literacy but not on numeracy scores. From Equations 7.12 and 7.13, it can be noted that the variable TRANS also has a significant influence on both numeracy and literacy scores.

Four things are worth noting for the analyses undertaken at this stage. First, without including the Prior Achievement variable in Model-Y and Z, the variable HOME (Speaking English at Home) also has a significant influence in numeracy. Thus, it is reasonable to assume that this variable has a significant influence on achievement in numeracy in Model-X because a prior achievement variable is not included in this type of model (see Figure 5.2).

Second, in all the analyses undertaken at this stage, the variable HOME is found to be a better predictor (that is, has a higher t-ratio value) than the alternative variable NESB (Non-English Speaking Background), and therefore, is chosen for inclusion in the models. However, where applicable, either of the two variables can be used since they both show a significant influence on achievement in numeracy and literacy if included in the model one at a time.

Third, no student-level variable has its effects specified as fixed at the school-level; that is, the effects associated with the Level-1 variables are left to vary across all the schools. And finally, the dummy variables SEX, TRANS, ATSI and YEARLEVL are entered into the equations uncentred but scaled variables AGE, HOME, INOZ and Y3NSCORE (or Y3LSCORE) are entered into the equations grand-mean-centred.

The results of the two-level HLM analyses using Equations 7.11 to 7.15 presented above, provide the final estimations of the fixed effects for each variable in the equation, the final estimations of the variance components and the deviance statistics of the unconditional models. The results also provide the reliability estimates[16] at Level-1 of the model for each variable with random effects at that level.

[16] For the interested reader, these reliability estimates are presented in Appendix 14.3.

The final estimations of the fixed effects for these unconditional models are presented in Tables 7.5 and 7.6 for numeracy and literacy respectively. Both the standardized as well as the metric regression coefficients of the variables in the final unconditional models are presented in Tables 7.5 and 7.6. The metric regression coefficients are obtained from HLM runs using raw scores of the variables while the standardized regression coefficients are obtained from separate HLM runs using standardized scores of the variables. The standardization of these variables was carried out using the *SPSS 10.0.5 for Windows* software.

The sizes of standardized coefficients of the variables indicate the relative magnitude of effects and can therefore be used to rank the variables in terms of their relative degree of influence on the outcome. However, the sizes of metric coefficients of the variables do not indicate the relative magnitude of effects and can not therefore be used to compare the degree of influence of the variables on the outcome (Hox, 1995).

For Model-X, the results in Tables 7.5 and 7.6 indicate that the Grade Level has a significant ($p<0.05$) influence on achievement in both numeracy and literacy even after taking into account other student-level variables. In addition, the metric growth coefficients recorded in Tables 7.5 and 7.6 for numeracy and literacy (0.52, 0.55) are almost equal to those obtained in the grade-level-only model (results in Table 7.2). These results indicate that the inclusion of student background variables into the regression equations do not substantially affect the values for the growth rates. Moreover, the standardized regression coefficients indicate that Grade Level (YEARLEVL) has the greatest magnitude of effect on achievement in numeracy (0.52) and literacy (0.55) compared with the other three student-level variables (AGE, ATSI and HOME) in the model.

For Models Y and Z, the results in Tables 7.5 and 7.6 indicate that of all the student-level variables examined, Prior Achievement (that is, achievement at Grade 3) in numeracy and literacy has the greatest magnitude of effect on achievement in numeracy and literacy at Grade 5.

Final two-level models

Further HLM runs are undertaken to build up the equations at the school-level through adding the significant school-level variables to the equation using the step-up strategy mentioned above. In this stage, an exploratory analysis sub-routine available in HLM5/2L is employed for examining the inclusion of potentially significant Level-2 predictors (as shown in the output) in successive HLM runs. This sub-routine allows for a maximum of 12 Level-2 predictors to be examined at a time for each variable at Level-1 that is specified as having a random effect at Level-2.

Table 7.7 presents examples of the results of the Level-2 exploratory analysis undertaken at the conclusion of the unconditional model HLM run for numeracy in Model-X. Similar exploratory analysis results are also obtained for numeracy and literacy in Models Y and Z following the same sub-routine. A step-by-step procedure is followed to select one potential predictor at a time to be added to the equation in the next HLM run. This is achieved by selecting the predictor with the highest t-value from the exploratory analysis results.

Table 7.5 Final estimation of fixed effects from the two-level unconditional models for numeracy

		Model - X					Model - Y					Model - Z				
		Coefficient[ξ]		SE	T-ratio	P-value	Coefficient[ξ]		SE	T-ratio	P-value	Coefficient[ξ]		SE	T-ratio	P-value
		Std'zed	Metric				Std'zed	Metric				Std'zed	Metric			
For **INTRCPT1**, B0																
	INTRCPT2, G00	0.72	-0.39	0.02	-17.53	0.00	1.37	1.20	0.03	41.83	0.00	1.42	1.24	0.03	39.60	0.00
For **SEX**, B1																
	INTRCPT2, G10	×××	×××	×××	×××	×××	-0.04	-0.09	0.01	-10.44	0.00	-0.05	-0.10	0.01	-10.94	0.00
For **AGE**, B2																
	INTRCPT2, G20	-0.11	-0.29	0.01	-33.29	0.00	-0.07	-0.20	0.01	-16.09	0.00	-0.07	-0.19	0.01	-14.15	0.00
For **ATSI**, B3																
	INTRCPT2, G30	0.12	0.61	0.02	32.26	0.00	0.04	0.24	0.03	8.76	0.00	0.04	0.24	0.03	8.08	0.00
For **HOME**, B4																
	INTRCPT2, G40	0.09	0.14	0.01	25.67	0.00	×××	×××	×××	×××	×××	×××	×××	×××	×××	×××
For **INOZ**, B5																
	INTRCPT2, G50	×××	×××	×××	×××	×××	-0.01	-0.04	0.01	-3.45	0.00	-0.01	-0.04	0.01	-2.93	0.00
For **TRANS**, B6																
	INTRCPT2, G60						-0.04	-0.10	0.01	-7.63	0.00					
For **YEARLEVL**, B7																
	INTRCPT2, G70	0.52	0.52	0.01	104.47	0.00										
For **Y3NSCORE**, B8																
	INTRCPT2, G80						0.70	0.56	0.01	124.76	0.00	0.71	0.58	0.01	119.07	0.00

Notes: ××× - Variable has no significant influence on the outcome
Shade - Variable not available for examination in this model
ξ - The standard errors (SE), t-ratios and p-values presented are those obtained using unstandardized variables.

Table 7.6 Final estimation of fixed effects from the two-level unconditional models for literacy

		Model - X					Model - Y					Model - Z				
		Coefficient[ξ]		SE	T-ratio	P-value	Coefficient[ξ]		SE	T-ratio	P-value	Coefficient[ξ]		SE	T-ratio	P-value
		Std'zed	Metric				Std'zed	Metric				Std'zed	Metric			
For	**INTRCPT1**, B0															
	INTRCPT2, G00	0.80	-0.36	0.02	-16.27	0.00	1.49	1.29	0.03	48.95	0.00	1.53	1.34	0.03	46.16	0.00
For	**SEX**, B1															
	INTRCPT2, G10	×××	×××	×××	×××	×××	0.02	0.04	0.01	4.86	0.00	0.02	0.03	0.01	3.66	0.00
For	**AGE**, B2															
	INTRCPT2, G20	-0.14	-0.37	0.01	-42.33	0.00	-0.06	-0.18	0.01	-16.28	0.00	-0.06	-0.17	0.01	-14.89	0.00
For	**ATSI**, B3															
	INTRCPT2, G30	0.12	0.62	0.02	31.84	0.00	0.04	0.19	0.03	7.41	0.00	0.03	0.18	0.03	6.42	0.00
For	**HOME**, B4															
	INTRCPT2, G40	0.09	0.15	0.01	29.06	0.00	0.02	0.04	0.01	4.05	0.00	0.01	0.03	0.01	3.20	0.00
For	**INOZ**, B5															
	INTRCPT2, G50	×××	×××	×××	×××	×××	-0.02	-0.05	0.01	-4.82	0.00	-0.02	-0.06	0.01	-4.66	0.00
For	**TRANS**, B6															
	INTRCPT2, G60						-0.02	-0.06	0.01	-4.94	0.00					
For	**YEARLEVL**, B7															
	INTRCPT2, G70	0.55	0.55	0.01	120.20	0.00										
For	**Y3LSCORE**, B8															
	INTRCPT2, G80						0.78	0.62	0.00	157.58	0.00	0.79	0.63	0.00	150.75	0.00

Notes: ××× - Variable has no significant influence on the outcome
Shade - Variable not available for examination in this model
ξ - The standard errors (SE), t-ratios and p-values presented are those obtained using unstandardized variables.

Table 7.7 Results of Level-2 exploratory analysis for Model-X numeracy

Level-1 Coefficient	**Potential Level-2 Predictors**					
	SEX_1	AGE_1	ATSI_1	HOME_1	INOZ_1	PSCARD
INTRCPT1,B0						
Coefficient	0.095	-0.052	0.117	0.482	-0.129	-1.413
Standard Error	0.084	0.113	0.212	0.032	0.071	0.049
t value	1.135	-0.464	0.551	15.280	-1.807	-29.057
	METRO	MOBILITY	CAP	ABSENT	SSIZELOG	GPOLOG
INTRCPT1,B0						
Coefficient	0.095	-0.011	-0.023	-4.101	0.041	-0.160
Standard Error	0.021	0.001	0.024	0.198	0.027	0.015
t value	4.481	-20.536	-0.945	-20.674	1.544	-10.399
	SEX_1	AGE_1	ATSI_1	HOME_1	INOZ_1	PSCARD
YEARLEVL,B1						
Coefficient	0.011	0.031	-0.052	-0.009	-0.020	0.042
Standard Error	0.020	0.027	0.051	0.008	0.017	0.013
t value	0.521	1.136	-1.010	-1.096	-1.182	3.103
	METRO	MOBILITY	CAP	ABSENT	SSIZELOG	GPOLOG
YEARLEVL,B1						
Coefficient	0.003	0.000	-0.017	0.029	0.020	-0.003
Standard Error	0.005	0.000	0.006	0.052	0.007	0.004
t value	0.664	1.475	-2.934	0.558	3.017	-0.720
	SEX_1	ATSI_1	HOME_1	INOZ_1	PSCARD	METRO
AGE,B2						
Coefficient	-0.001	-0.008	0.013	0.009	-0.032	0.001
Standard Error	0.011	0.027	0.004	0.009	0.007	0.003
t value	-0.090	-0.289	3.193	0.994	-4.607	0.228
	MOBILITY	CAP	ABSENT	SSIZELOG	GPOLOG	YR35PPT
AGE,B2						
Coefficient	-0.000	0.005	-0.070	0.004	0.001	0.000
Standard Error	0.000	0.003	0.027	0.003	0.002	0.000
t value	-3.852	1.698	-2.604	1.045	0.691	1.274
	SEX_1	AGE_1	HOME_1	INOZ_1	PSCARD	METRO
ATSI,B3						
Coefficient	-0.038	0.015	-0.164	0.065	0.274	-0.051
Standard Error	0.023	0.031	0.009	0.020	0.015	0.006
t value	-1.620	0.488	-19.008	3.256	18.892	-8.631
	MOBILITY	CAP	ABSENT	SSIZELOG	GPOLOG	YR35PPT
ATSI,B3						
Coefficient	0.003	0.053	1.421	-0.043	0.058	-0.001
Standard Error	0.000	0.007	0.053	0.007	0.004	0.000
t value	23.527	7.808	26.844	-5.811	13.570	-10.419
	SEX_1	AGE_1	ATSI_1	INOZ_1	PSCARD	METRO
HOME,B4						
Coefficient	-0.010	-0.002	-0.005	0.031	0.030	-0.012
Standard Error	0.006	0.008	0.015	0.005	0.004	0.002
t value	-1.625	-0.300	-0.334	5.893	7.480	-8.119
	MOBILITY	CAP	ABSENT	SSIZELOG	GPOLOG	YR35PPT
HOME,B4						
Coefficient	0.001	0.015	0.250	-0.010	0.013	-0.000
Standard Error	0.000	0.002	0.015	0.002	0.001	0.000
t value	13.475	8.342	16.849	-5.379	11.859	-7.274

For example, for the results of numeracy in Model-X presented in Table 7.7, PSCARD at Level-2 associated with INTRCPT1 has the highest 't-to-enter' value (-29.06) so this is the predictor that is added to the equation in the next HLM run. In

addition, the exploratory analysis provides an estimate of what the coefficient of the potential predictor would be if added to the equation. For numeracy in the Model-X example, the coefficient of PSCARD when added at Level-2 of the model, will be approximately -1.41, indicating that schools with a higher proportion of school cardholders, are likely to have a lower intercept than schools with a lower proportion of school cardholders.

However, it should be noted that the t-ratio value in the exploratory analysis represents the approximate result that would be obtained when one additional predictor is added to any of the Level-2 equations (Raudenbush et al., 2000). This means that, for example, when PSCARD is added to the model for the INTRCPT1 the apparent relationship suggested in Table 7.7 for MOBILITY, in the ATSI slope model disappears. Some potential predictors, after being included in the equation, provide absolute t-ratio values that are smaller than 1.96 in the final estimation of the fixed effects in the two-level HLM output. Consequently, these predictors are deleted from further analysis.

The final two-level hierarchical models for numeracy and literacy for the three types of models specified are presented in Figures 7.1 to 7.6.

The final estimations of the fixed effects from the two-level HLM analysis for the models shown in Figures 7.1 to 7.6 are displayed in Tables 7.8 and 7.9 for numeracy and literacy respectively. Both the standardized as well as the metric regression coefficients of the variables in the final models are presented in the two tables.

In the next three sub-sections, discussion of the results of the final fixed effects displayed in Tables 7.8 and 7.9 are presented. A discussion of the final models at the student-level is presented, followed by a discussion of the models at the school-level for the three types of models examined, and then a summary of the interaction effects in these models. A thorough treatment of these interactions can be found in Hungi (2003; pp.503-528).

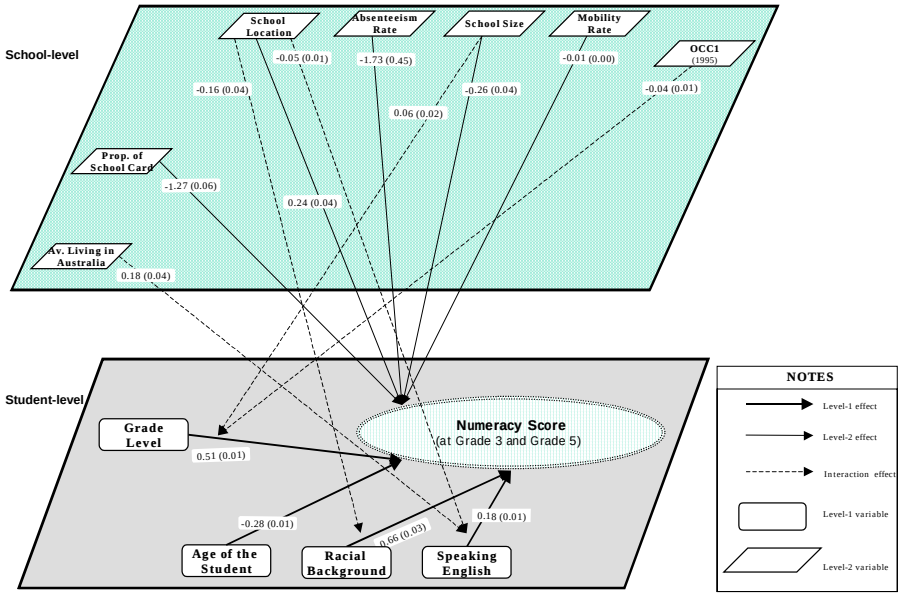


Figure 7.1 Final two-level hierarchical model for numeracy - Model-X

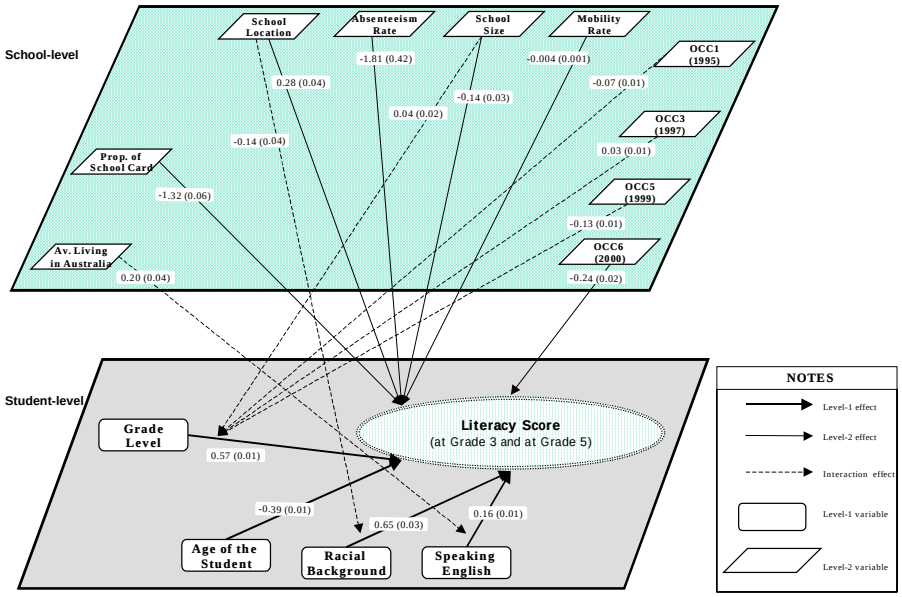


Figure 7.2 Final two-level hierarchical model for literacy - Model-X

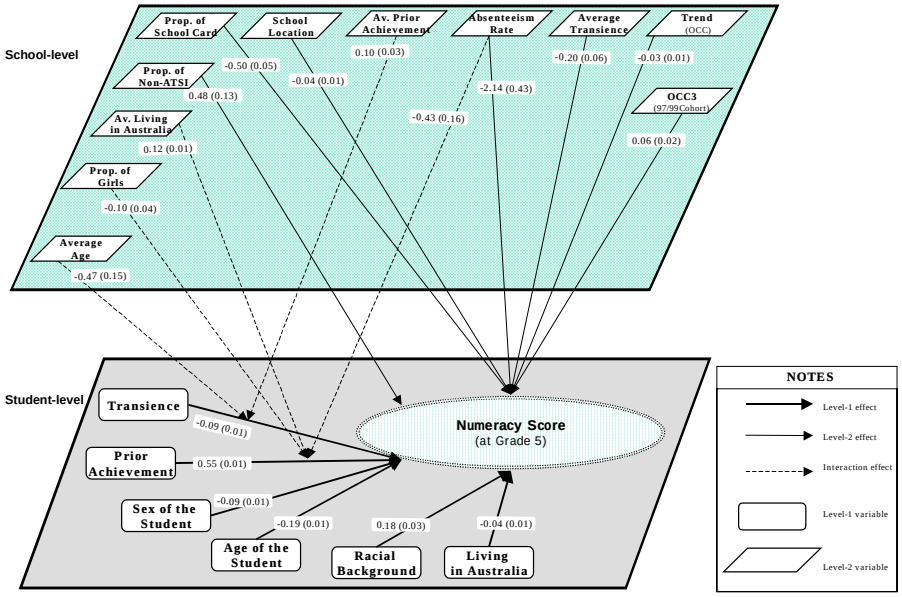


Figure 7.3 Final two-level hierarchical model for numeracy - Model-Y

Student-level model

For Model-X, the results in Tables 7.8 and 7.9 indicate that there are four student-level variables that have a significant ($p<0.05$) influence on achievement in numeracy, namely YEARLEVL (Grade-level), AGE, ATSI (Racial Background) and HOME

(Speaking English at Home). These four student-level variables also have a significant influence on achievement in literacy. The standardized coefficients of these four student-level variables indicate that Grade Level[17] (0.51, 0.55) has by far the greatest magnitude of effect, followed by the Age of the Student (-0.11, -0.14) and the Racial Background of the Student (0.11, 0.11) which have almost the equal magnitudes of effects. Of these four variables, the Speaking English at Home (0.10, 0.10) has the lowest magnitude of effect.

For both numeracy and literacy, the coefficients and the t-ratio values for YEARLEVL are positive indicating that students at Grade 5 are likely to perform better in numeracy and literacy than students at Grade 3. Consequently, as shown by the metric coefficients, the growth in numeracy achievement is 0.51 logits per year of study, and for literacy, it is 0.57 logits per year of study. These growths in numeracy and literacy achievement between the two grades have remained almost the same as those obtained from the grade-level-only models (results in Table 7.2) and the unconditional models (results in Tables 7.5 and 7.6). However, it should be noted from the results in Tables 7.8 and 7.9 that there are interaction effects between YEARLEVL with the dummy variables for some occasions indicating that the estimated average growth in achievement varied significantly for some testing occasions.

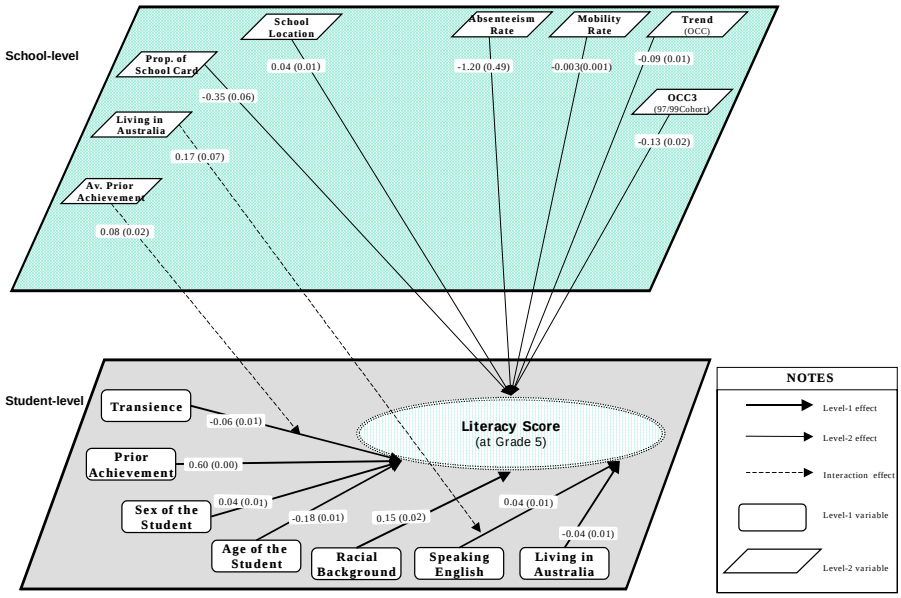


Figure 7.4 Final two-level hierarchical model for literacy - Model-Y

The negative coefficients and the negative t-ratio values for the interactions between YEARLEVL and OCC1 (-0.04, t=-4.09 for numeracy; -0.07, t=-6.36 for literacy) indicate that in 1995 the growths in achievement in the two subjects is estimated to be lower than on the other testing occasions. Likewise, in 1999 (OCC5) the growth in achievement in literacy (-0.13, t=-14.02) is estimated to be lower than on the other

[17] In this paragraph, the first coefficient listed in parenthesis is for numeracy and the second value is for literacy.

occasions. On the other hand, the positive coefficient and positive t-ratio value for the interactions between YEARLEVL and OCC3 for literacy (0.03, t=2.74) indicate that in 1997 the growth in literacy is estimated to be higher than on the other occasions.

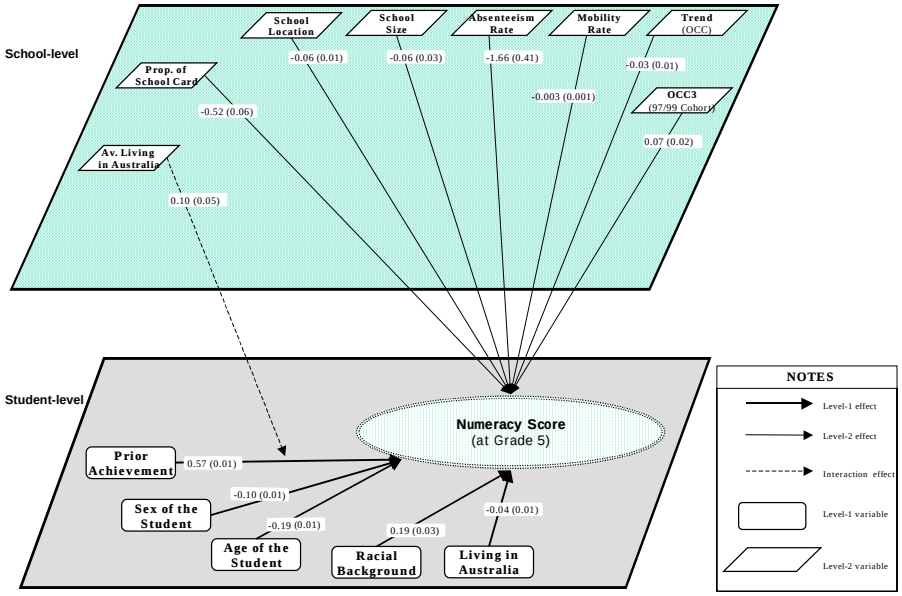


Figure 7.5 Final two-level hierarchical model for numeracy - Model-Z

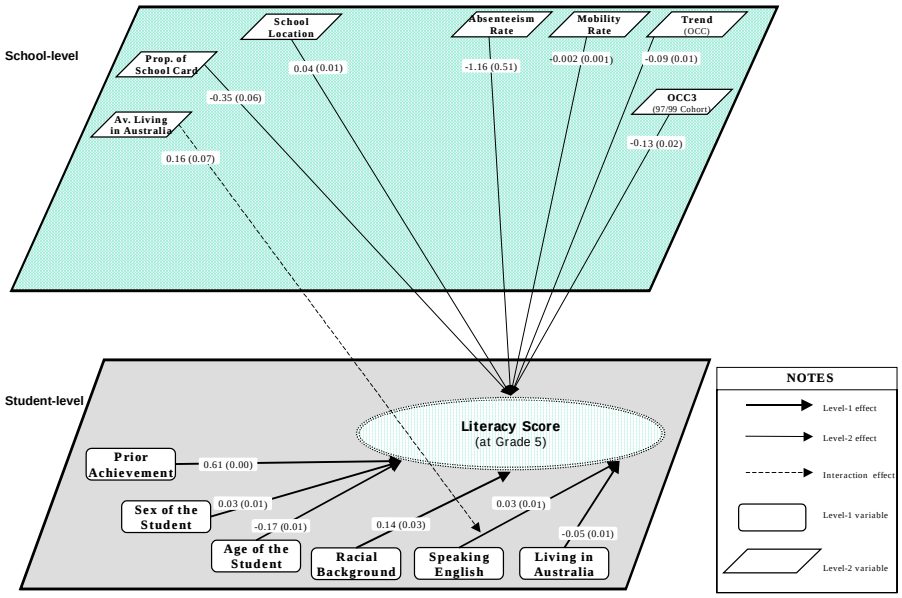


Figure 7.6 Final two-level hierarchical model for literacy - Model-Z

Table 7.8 Final estimation of fixed effects from the final two-level numeracy models

			Model-X					Model-Y					Model-Z				
			Coefficient[ξ]		SE	T-ratio	P-value	Coefficient[ξ]		SE	T-ratio	P-value	Coefficient[ξ]		SE	T-ratio	P-value
			Std'zed	Metric				Std'zed	Metric				Std'zed	Metric			
For	**INTRCPT1**,	B0															
	INTRCPT2, G00		0.70	-0.49	0.03	-17.10	0.00	1.34	1.21	0.03	45.69	0.00	1.40	1.24	0.03	41.17	0.00
	METRO, G01		0.04	0.24	0.04	6.29	0.00										
	PSCARD, G02		-0.25	-1.27	0.06	-22.12	0.00	-0.10	-0.50	0.05	-9.98	0.00	-0.10	-0.52	0.06	-9.15	0.00
	MOBILITY, G03		-0.09	-0.01	0.00	-4.72	0.00						-0.05	0.00	0.00	-2.79	0.01
	ABSENT, G04		-0.09	-1.73	0.45	-3.88	0.00	-0.10	-2.14	0.43	-5.01	0.00	-0.08	-1.66	0.41	-4.06	0.00
	SSIZELOG, G05		-0.08	-0.26	0.04	-7.37	0.00						-0.02	-0.06	0.03	-1.98	0.05
	GPOLOG, G06							-0.03	-0.04	0.01	-3.51	0.00	-0.04	-0.06	0.01	-4.53	0.00
	ATSI_1, G07							0.08	0.48	0.13	3.61	0.00					
	TRANS_1, G08							-0.04	-0.20	0.06	-3.34	0.00					
	OCC3, G09							0.03	0.06	0.02	3.46	0.00	0.03	0.07	0.02	3.39	0.00
	OCC, G010							-0.03	-0.03	0.01	-4.07	0.00	-0.03	-0.03	0.01	-4.21	0.00
For	**SEX**,	B1															
	INTRCPT2, G10		×××	×××	×××	×××	×××	-0.05	-0.09	0.01	-10.96	0.00 s	-0.05	-0.10	0.01	-11.02	0.00 s
For	**AGE**,	B2															
	INTRCPT2, G20		-0.11	-0.28	0.01	-32.53	0.00 s	-0.07	-0.19	0.01	-15.21	0.00 s	-0.07	-0.19	0.01	-13.52	0.00 s
For	**ATSI**,	B3															
	INTRCPT2, G30		0.11	0.66	0.03	24.75	0.00	0.03	0.18	0.03	7.01	0.00 s	0.03	0.19	0.03	6.76	0.00 s
	METRO, G31		-0.02	-0.16	0.04	-4.46	0.00										
For	**HOME**,	B4															
	INTRCPT2, G40		0.10	0.18	0.01	16.42	0.00	×××	×××	×××	×××	×××	×××	×××	×××	×××	×××
	INOZ_1, G41		0.02	0.18	0.04	4.59	0.00										
	METRO, G42		-0.02	-0.05	0.01	-3.73	0.00										
For	**INOZ**,	B5															
	INTRCPT2, G50		×××	×××	×××	×××	×××	-0.01	-0.04	0.01	-3.06	0.00 s	-0.01	-0.04	0.01	-2.98	0.00 s
For	**TRANS**,	B6															
	INTRCPT2, G60							-0.03	-0.09	0.01	-7.01	0.00					
	Y3NSCO_1, G61							0.03	0.10	0.03	4.04	0.00					
	AGE_1, G62							-0.03	-0.47	0.15	-3.15	0.00					
For	**YEARLEVL**,	B7															
	INTRCPT2, G70		0.51	0.51	0.01	85.93	0.00										
	SSIZELOG, G71		0.02	0.06	0.02	3.51	0.00										
	OCC1, G72		-0.03	-0.04	0.01	-4.09	0.00										
For	**Y3NSCORE**,	B8															
	INTRCPT2, G80							0.69	0.55	0.01	118.63	0.00	0.70	0.57	0.01	115.51	0.00
	ABSENT, G81							-0.03	-0.43	0.16	-2.65	0.01					
	SEX_1, G82							-0.02	-0.10	0.04	-2.73	0.01					
	INOZ_1, G83							0.02	0.12	0.05	2.56	0.01	0.01	0.10	0.05	2.13	0.03

Notes: ××× - Variable has no significant influence on the outcome.
Shade - Variable not available for examination in this model.
ξ - Standard errors (SE), t-ratios and p-values presented are those obtained using unstandardized variables.
s - Residual parameter of this coefficient is fixed at the school-level.

Table 7.9 Final estimation of fixed effects from the final two-level literacy models

			Model-X					Model-Y					Model-Z				
			Coefficient[ξ]		SE	T-ratio	P-value	Coefficient[ξ]		SE	T-ratio	P-value	Coefficient[ξ]		SE	T-ratio	P-value
			Std'zed	Metric				Std'zed	Metric				Std'zed	Metric			
For	**INTRCPT1**,	B0															
	INTRCPT2, G00		0.77	-0.44	0.03	-14.44	0.00	1.47	1.32	0.03	51.73	0.00	1.51	1.37	0.03	48.41	0.00
	METRO, G01		0.07	0.28	0.04	7.20	0.00	0.02	0.04	0.01	3.17	0.00	0.02	0.04	0.01	2.89	0.00
	PSCARD, G02		-0.25	-1.32	0.06	-24.14	0.00	-0.07	-0.35	0.06	-5.91	0.00	-0.06	-0.35	0.06	-5.52	0.00
	MOBILITY, G03		-0.07	0.00	0.00	-3.74	0.00	-0.05	0.00	0.00	-2.61	0.01	-0.04	0.00	0.00	-1.99	0.05
	ABSENT, G04		-0.09	-1.81	0.42	-4.37	0.00	-0.06	-1.20	0.49	-2.45	0.01	-0.05	-1.16	0.51	-2.27	0.02
	SSIZELOG, G05		-0.04	-0.14	0.03	-4.47	0.00										
	OCC3, G06							-0.05	-0.13	0.02	-8.41	0.00	-0.06	-0.13	0.02	-8.34	0.00
	OCC6, G07		-0.08	-0.24	0.02	-13.65	0.00										
	OCC, G08							-0.10	-0.09	0.01	-13.08	0.00	-0.10	-0.09	0.01	-12.69	0.00
For	**SEX**,	B1															
	INTRCPT2, G10		×××	×××	×××	×××	×××	0.02	0.04	0.01	4.94	0.00 *s*	0.02	0.03	0.01	3.79	0.00 *s*
For	**AGE**,	B2															
	INTRCPT2, G20		-0.14	-0.36	0.01	-41.63	0.00 *s*	-0.06	-0.18	0.01	-16.16	0.00 *s*	-0.06	-0.17	0.01	-14.81	0.00 *s*
For	**ATSI**,	B3															
	INTRCPT2, G30		0.11	0.65	0.03	22.75	0.00	0.03	0.15	0.02	6.45	0.00 *s*	0.03	0.14	0.03	5.49	0.00 *s*
	METRO, G31		-0.01	-0.14	0.04	-3.94	0.00										
For	**HOME**,	B4															
	INTRCPT2, G40		0.10	0.16	0.01	27.79	0.00	0.02	0.04	0.01	3.77	0.00	0.02	0.03	0.01	3.26	0.00
	INOZ_1, G41		0.02	0.20	0.04	5.15	0.00	0.01	0.17	0.07	2.39	0.02	0.01	0.16	0.07	2.24	0.03
For	**INOZ**,	B5															
	INTRCPT2, G50		×××	×××	×××	×××	×××	-0.02	-0.04	0.01	-4.14	0.00 *s*	-0.02	-0.05	0.01	-3.94	0.00 *s*
For	**TRANS**,	B6															
	INTRCPT2, G60							-0.02	-0.06	0.01	-4.99	0.00					
	Y3LSCO_1, G61							0.02	0.08	0.02	3.12	0.00					
For	**YEARLEVL**,	B7															
	INTRCPT2, G70		0.55	0.57	0.01	88.16	0.00										
	SSIZELOG, G71		0.02	0.04	0.02	2.52	0.01										
	OCC1, G72		-0.04	-0.07	0.01	-6.36	0.00										
	OCC3, G73		0.01	0.03	0.01	2.74	0.01										
	OCC5, G74		-0.05	-0.13	0.01	-14.02	0.00										
For	**Y3LSCORE**,	B8															
	INTRCPT2, G80							0.77	0.60	0.00	152.95	0.00 *s*	0.77	0.61	0.00	144.79	0.00 *s*

Notes: ××× - Variable has no significant influence on the outcome.
Shade - Variable not available for examination in this model.
ξ - Standard errors (SE), t-ratios and p-values presented are those obtained using unstandardized variables.
s - Residual parameter of this coefficient is fixed at the school-level.

For AGE, the negative coefficients and t-ratio values indicate that students of younger age are likely to achieve better in numeracy (-0.28, t=-32.53) and literacy (-0.36, t=-41.63) than their older counterparts. For ATSI, the positive values indicate that Aboriginal and Torres Strait Islander students (coded as ATSI=0) are likely to perform at a lower level than students of other racial backgrounds (coded as ATSI=1) in numeracy (0.66, t=24.75) and also in literacy (0.65, t=22.75). Finally, the positive values for HOME indicate that students who always speak English at home (coded as HOME=3) are likely to achieve better than students who never speak English at home (coded as HOME=0) in numeracy (0.18, t=16.42) and in literacy (0.16, t=27.79).

For Model-Y, the results in Tables 7.8 and 7.9 indicate that there are six student-level variables that have a significant influence on achievement in both numeracy and literacy, namely SEX, AGE ATSI, INOZ, TRANS and Y3NSCORE (or Y3LSCORE for literacy). In addition, Model-Y results presented in the two tables indicate that at the student-level the variable HOME (Speaking English at Home) has a significant influence on achievement in literacy (0.04, t=3.77) but not in numeracy.

From the results presented in Tables 7.8 and 7.9 it can be observed that, apart from the variable TRANS, the same student-level variables that have a significant influence on achievement in either numeracy or literacy using Model-Y correspondingly have a significant influence in Model-Z. However, the variable TRANS is not available for testing in Model-Z since the data used in this type of model are from students who did not change schools between the two grades (see Chapter 5.2).

The values of the coefficients and the t-ratios for Prior Achievement (Y3NSCORE or Y3LSCORE) are positive indicating that those students who had higher performance on the tests while at Grade 3, are likely to achieve better on the Grade 5 tests than their counterparts who had lower performance. It should be observed that the sizes of the metric coefficients for Prior Achievements recorded here have not changed considerably compared to those obtained using the prior-achievement-only models (results in Table 7.3) and the unconditional models (results in Tables 7.5 and 7.6). Hence, it appears that the inclusion of other variables into the models does not change greatly the effect sizes associated with Prior Achievement. Furthermore, the standardized coefficients provided in Tables 7.8 and 7.9 indicate that in Models Y and Z, Prior Achievement has by far the greatest magnitude of effect for both numeracy (0.70) and literacy (0.77).

Similar to what is recorded from Model-X, the results in Tables 7.8 and 7.9 for Models Y and Z indicate that for both subjects (a) students of younger age are likely to achieve better than their older counterparts, and (b) ATSI students are likely to perform at a lower level than non-ATSI students. The results for Models Y and Z also indicate that students who always speak English at home are likely to achieve better in literacy than students who never speak English at home. However, in contrast with what is recorded using Model-X, these two types of models indicate that English spoken in the home has no significant influence on achievement in numeracy.

It has been mentioned above that using Model-X the Gender of the Student (SEX) and Migrant Status in Australia (INOZ) are not found to have significant influences on achievement in either numeracy or literacy. However, the results for Models Y and Z indicate that SEX and INOZ have significant influences on achievement in both numeracy and literacy. For numeracy as well as literacy, the coefficient and t-ratio values for INOZ are negative indicating that Grade 5 students who have lived a shorter time in Australia are likely to achieve better than their counterparts who have lived in Australia. However, for the variable SEX, the coefficient and the t-ratio values for numeracy are negative while the values are positive for literacy. For

numeracy, the negative coefficients and t-ratio values for SEX indicate that boys (coded SEX=0) are likely to achieve better in numeracy than girls (coded SEX=1). On the other hand, for literacy, the positive coefficients and t-ratio values for SEX indicate that girls are likely to achieve better in literacy than boys.

Finally, results in Tables 7.8 and 7.9 for Model-Y indicate that the variable TRANS has a significant influence on achievement in both numeracy and literacy. In order to interpret appropriately these results it should be remembered that the predictor TRANS is coded (same school = 0, changed school = 1). Hence, the negative coefficients and negative t-values for TRANS indicate that Grade 5 students who remained in the same school that they were in at Grade 3 are likely to achieve better than their Grade 5 counterparts who changed schools in between the two grades. The metric coefficients for the variable TRANS provided in the two tables shows a significant average poorer performance for students who changed school of -0.09 logits (with a t-value of -7.01) for numeracy, and of -0.06 logits (with a t-value of -4.99) for literacy. These values follow closely those obtained from the unconditional models (results in Tables 7.5 and 7.6).

The estimated decline in achievement due to transience should be considered substantial because based on the estimated YEARLEVL effects, the increase in numeracy and literacy between the two grades is on average about one logit. Hence, the -0.09 logits indicate that the decline in achievement in numeracy due to transience is on average about three-quarters of a term (or eight weeks) of school learning. Similarly, the -0.06 logits indicate that the decline in achievement in literacy is on average about half a term (or six weeks) of school learning.

In summary, the results at the student-level for the three types of models agree, and they generally indicate the following relationships regarding the student-level factors influencing achievement in the BST among Grades 3 and 5 students in South Australia when other variables are equal.

1. **Grade Level**: Students at Grade 5 are likely to achieve better in numeracy and literacy than students at Grade 3 by about one logit, and therefore, the increase in achievement in numeracy and literacy between Grade 3 level and Grade 5 level is on average about one logit.
2. **Age of the Student**: Younger students are likely to achieve better in numeracy and literacy than their older counterparts.
3. **Racial Background**: Aboriginal and Torres Strait Islander students are likely to achieve lower in numeracy and literacy than other students.
4. **Speaking English at Home**: Students who always speak English at home are likely to achieve better in numeracy (Model-X only) and literacy (all the Models) than students who rarely (or never) speak English at home.
5. **Sex of the Student**: Boys are likely to achieve better in numeracy than girls, while girls are likely to achieve better in literacy than boys.
6. **Living in Australia**: Students who are new to Australia are likely to achieve better in numeracy and literacy than the students who were born in Australia.
7. **Transience**: Students who remain in the same school over the two-year duration are likely to achieve better in numeracy and literacy than students who change schools.
8. **Prior Achievement**: Students who have high scores in the BST at Grade 3 are likely to perform better on the tests at Grade 5 than students who have low scores on the tests at Grade 3.

School-level model

For Model-X, the results in Tables 7.8 and 7.9 indicate that five variables, namely PSCARD (Proportion of School-card[18] Holders), METRO (School Location), MOBILITY (Mobility Rate), ABSENT (Absenteeism Rate) and SSIZELOG (School Size), have significant influences on student achievement in numeracy and literacy at the school-level. In addition, the dummy variable OCC6 (2000) has a significant influence on achievement in literacy but not in numeracy. The sizes of the standardized coefficients of the variables provided in the two tables indicate the relative magnitude of effects. Thus, of the five variables with significant influences on achievement in numeracy and literacy, PSCARD has the largest magnitude of effect. On the other hand, the signs of the regression coefficients of these variables indicate the directions of effects and can only be interpreted appropriately if the coding of the variables is considered. Thus, the negative coefficients and the negative t-ratios for the variables PSCARD, MOBILITY, ABSENT, SSIZELOG in Model-X indicate the following effects on achievement in numeracy and literacy when other variables are equal.

Students in schools with lower proportions of school-card holders are likely to achieve better than their counterparts in schools with higher proportions of school-card holders while students in schools with lower mobility rates are likely to achieve better than students in schools with higher mobility rates. Students in schools with lower rates of absenteeism are likely to achieve better if they are compared to students in schools with higher rates of absenteeism. In addition, students in schools with fewer students are likely to achieve better than students in schools with many students.

The positive coefficients and positive t-ratio values for METRO (coded as urban=1, rural=0) in Tables 7.8 and 7.9 (for Model-X) indicate that, when other variables are equal, students in urban schools are likely to achieve better in numeracy and literacy than students in rural schools. In addition, these results for Model-X indicate that the performance of the students in 2000 (OCC6) in literacy (but not necessarily in numeracy) was likely to be lower than the performance of the students on the other five testing occasions.

Information displayed in Tables 7.8 and 7.9 show that in general the results for Models Y and Z are very similar to those described above for Model-X for both numeracy and literacy. However, there are a few differences between the results obtained from Model-X and the results obtained from the other two types of models at the school-level as outlined below.

First, the School Size variable (SSIZELOG) has no significant influence on achievement in literacy (for Models Y and Z), and has a significant influence on achievement in numeracy (for Model-Z only).

Second, for numeracy, the School Location variable (GPOLOG, Logarithm of Distance from Adelaide GPO) is found to be a better predictor (that is, has a higher t-ratio value) than its alternative variable METRO in Models Y and Z, and therefore is chosen for inclusion in the models. Consequently, the results here indicate that students in schools near Adelaide (the major urban centre in South Australia) are likely to achieve better in numeracy than students in schools that are far from Adelaide (that is, remotely located). However, it should be noted that for all three types of models examined either METRO or GPOLOG could be used since they both show a

[18] In South Australia, a school-card is provided to students of low social economic status so that they may obtain concessions in a number of services.

significant influence on achievement in numeracy (and also in literacy) when included in the models one at a time.

Third, for numeracy in Model-Y, the variable TRANS_1 (Proportion of Grade 5 Newcomers) is found to be a better predictor than its alternative variable MOBILITY (Mobility Rate) and is consequently chosen for inclusion in this model. Hence, the results here indicate that students in schools with low proportions of Grade 5 new students are likely to achieve better in numeracy than students in schools with high proportions of Grade 5 new students. However, it should also be noted that either MOBILITY or TRANS_1 could be used in Model-Y since they both show a significant influence on achievement in numeracy and literacy when included in the models one at a time.

Fourth, the results for numeracy Model-Y indicate that the variable ATSI_1 (Proportion of Grade 5 non-ATSI Students) has a significant influence on achievement in numeracy. The positive coefficient and t-ratio values for ATSI_1 (0.48, t=3.61) indicate that students in schools with a high proportion of non-ATSI students are likely to achieve better in numeracy than students in schools with a high proportion of Aboriginal and Torres Strait Islander students.

Finally, results for Models Y and Z indicate that the linear trend variable OCC and the dummy variable OCC3 (1997/1999 Cohort) have significant influences on achievement in both numeracy and literacy. The negative coefficients and t-ratio values for OCC indicate that, after allowance is made for other significant factors, there is a general decline in the performance of the schools between occasions. That is, the performance of the schools on the earlier occasions is estimated to be higher than their performance on the later occasions. For OCC3, the regression coefficients and the t-ratio values for numeracy are positive while these values are negative for literacy. For literacy, these results indicate that, when other variables are held equal, the performance of the schools based on the 1997/1999 Cohort of students is estimated to be lower than the performance of the schools based on the other three cohorts of students of interest in this study. And for numeracy, these results indicate that the performance of the schools based on the 1997/1999 Cohort is estimated to be higher than the performance of the schools based on the other three cohorts of students when other variables remain the same.

In summary, the results at the school-level for the three types of models generally agree and they mainly indicate the following relationships regarding the school-level factors influencing achievement in the BST among Grades 3 and 5 students in South Australia, when other variables are held equal.

1. **School-card** (Socioeconomic Status): Students in schools with lower proportions of school-card holders are likely to achieve better in numeracy and literacy than students in schools with higher proportions of school-card holders.
2. **School Location**: Students in schools located in urban areas (or near Adelaide) are likely to achieve better in numeracy and literacy than students in schools located in rural areas (or far from Adelaide).
3. **Mobility Rate**: Students in schools with low mobility rates are likely to achieve better in numeracy and literacy than students in schools with high mobility rates.
4. **Absenteeism Rate**: Students in schools with low rates of absenteeism are likely to achieve better in numeracy and literacy than students in schools with high rates of absenteeism.

5. **School Size**: Students in schools with fewer students are likely to achieve better in numeracy (Models X and Z) and literacy (Model-X only)than students in schools with many more students;
6. **Proportion of Aboriginal Students**: Students in schools with high proportions of non-ATSI students are likely to achieve better in numeracy (but not necessarily literacy) than students in schools with high proportions of Aboriginal and Torres Strait Islander students (Model-Y only).

Seven student-related variables examined at the school-level had no direct significant influence on achievement in numeracy and literacy in any of the three types of models examined. The seven variables are namely SEX_1, AGE_1, HOME_1, INOZ_1, CAP, YR35PPT and Y3NSCO_1 (for numeracy) or Y3LSCO_1 (for literacy).

Cross-level interaction effects

The results in Tables 7.8 and 7.9 also indicate the following nine interaction effects in Models X, Y and Z for numeracy and literacy between student-level variables and school-level variables.

1. Racial Background (ATSI) with School Location (METRO) in Model-X for numeracy and literacy.
2. Speaking English at Home (HOME) with School Location (METRO) in Model-X for numeracy.
3. Speaking English at Home (HOME) with Average Duration of Living in Australia (INOZ_1) in Models X and Y for numeracy and literacy.
4. Grade Level (YEARLEVL) with School Size (SSIZELOG) in Model-X for numeracy and literacy.
5. Transience (TRANS) with Average Prior Achievement (Y3NSCO_1 or Y3LSCO_1) in Model-Y for numeracy and literacy.
6. Transience (TRANS) with Average Age of the Students (AGE_1) in Model-Y for numeracy.
7. Prior Achievement (Y3NSCORE) with Absenteeism Rate in the School (ABSENT) in Model-Y for numeracy.
8. Prior Achievement (Y3NSCORE) with Proportion of Girls in the Schools (SEX_1) in Model-Y for numeracy.
9. Prior Achievement (Y3NSCORE) with Average Duration of Living in Australia (INOZ_1) in Models Y and Z for numeracy.

These interactions are cross-level interactions since they involve interaction between variables at a higher level with variables at a lower level (Raudenbush and Willms, 1991). Hox (1995, p.26) has pointed out that "the effect of the interaction and the direct effects of the explanatory variable that make up the interaction must be interpreted as a system".

In the paragraphs that follow, a summary of the interpretations of the cross-level interaction effects listed above is presented. Graphical representations of the interaction effects are used to enhance this summary. The coordinates of the graphs used in this summary (Figures 7.7 to 7.15) are calculated from the final estimation of the fixed effects obtained from the final models (results in Tables 7.8 and 7.9). Aiken and West (1996) and also Lietz (1996) have described the procedure employed to calculate the graphs for the interaction effects.

It should be noted that, where the same variables are involved, the graphical plots for the interaction effects for numeracy achievement are found to be basically identical to the corresponding plots for literacy in this study. Thus, in order to avoid repetition, only the graphs for numeracy have been presented here but it should be borne in mind that, where it is applicable, whatever conclusions reached for numeracy achievement applies for literacy achievement as well.

Thus, the results displayed in Tables 7.8 and 7.9, and the graphical plots in Figures 7.7 to 7.15 indicate the following relationships regarding the impact of the interaction effects between student-level factors with school-level factors on performance of the students in the Basic Skills Tests when other variables are equal.

1. **Racial Background with School Location**: The graphical representation in Figure 7.7 shows that non-ATSI students are estimated to achieve better in numeracy and literacy than ATSI students regardless of the school locality. However, the difference in achievement in numeracy (and in literacy) between ATSI students in rural and urban schools is estimated to be larger when compared to the corresponding difference between non-ATSI students in rural and urban schools. In other words, locality of the school has a greater influence on achievement in numeracy and literacy of ATSI students than of non-ATSI students, with ATSI students in rural schools being estimated to achieve at a much lower level than would be expected.
2. **Speaking English at Home with School Location**: In Figure 7.8 it can be seen that students who always speak English at home are estimated to achieve better in numeracy than students who rarely (or never) speak English at home regardless of the school locality. However, locality of the school has a greater impact on achievement in numeracy of students who rarely (or never) speak English at home than of students who always speak English at home. Consequently, students who rarely (or never) speak English at home are estimated to achieve at a much lower level in numeracy than would be expected if they were in schools located in rural areas compared with if they were in schools located in urban areas.

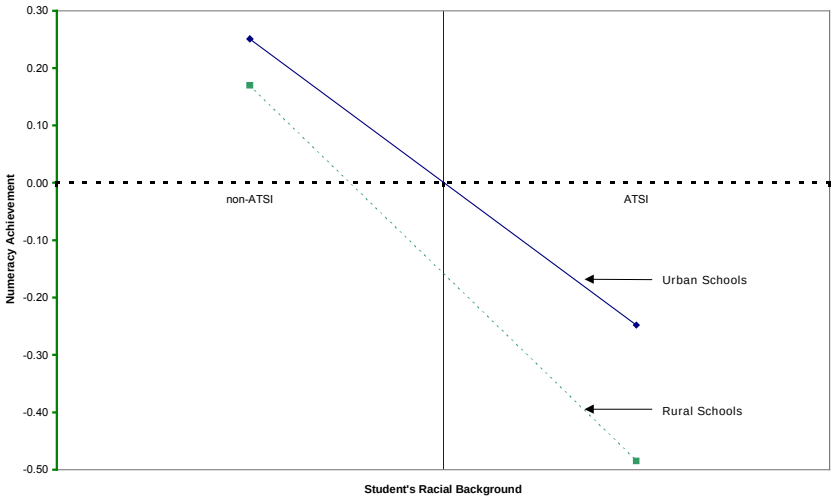


Figure 7.7 Impact of the interaction effect of student's Racial Background with Schools Location on Numeracy achievement

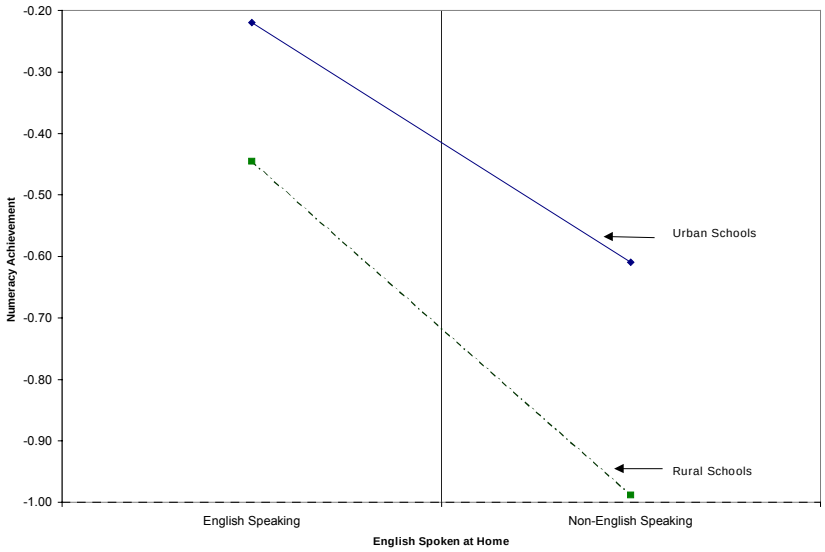


Figure 7.8 Impact of the interaction effect of Speaking English at Home with School Location on Numeracy achievement

3. **Speaking English at Home with Average Duration of Living in Australia**: The graph in Figure 7.9 shows that students who always speak English at home are estimated to perform equally well in numeracy regardless of whether they are in schools with large proportions of students born in Australia or they are schools with large proportions of students who are new to Australia. However, students who never speak English at home are estimated to perform better in numeracy if in schools with large proportions of students who have lived in Australia for a short duration than if in schools with large proportions of students born in Australia.
4. **Grade Level with School Size**: From Figure 7.10 it can be seen that at both Grade 3 and Grade 5 levels, students in schools with fewer pupils (small schools) are estimated to achieve better in numeracy and literacy than students in schools with many pupils (large schools). However, the difference in achievement between Grade 3 students in small schools and Grade 3 students in large schools is estimated to be significantly larger than the corresponding difference between Grade 5 students in small and large schools. That is, school size has a greater impact on the achievement in numeracy and literacy of Grade 3 students than of Grade 5 students, with Grade 3 students in large schools achieving much lower than would be expected.
5. **Transience with Average Prior Achievement**: The graph in Figure 7.11 shows that students who do not change schools between Grades 3 and 5 are estimated to achieve equally well in numeracy (or in literacy) at Grade 5 regardless of whether they are in schools that have high, average or low prior achievement scores. However, transient students who move to schools that have high prior achievement in numeracy (or literacy) are estimated to achieve better in numeracy (or literacy) than students who move to schools with average or low prior achievement in numeracy (or in literacy).

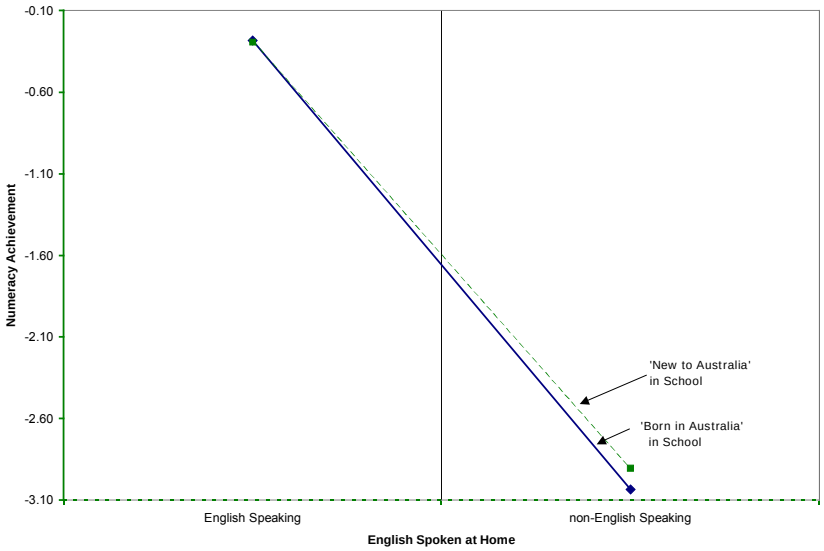


Figure 7.9 Impact of the interaction effect of student's Speaking English at Home with Average Living in Australia in schools on Numeracy achievement

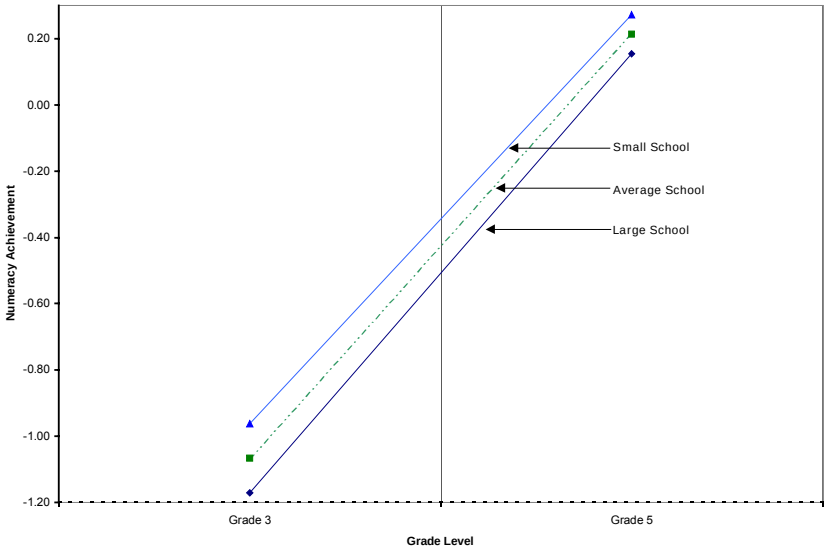


Figure 7.10 Impact of the interaction effect of Grade Level with School Size on Numeracy achievement

6. **Transience with Average Age of the Students**: It can be seen from Figure 7.12 that non-transient students are estimated to achieve equally well in numeracy (or in literacy) at Grade 5 regardless of whether they are in schools that have high proportions of older students or in schools with high proportions of younger students. Nevertheless, students who move into schools that have high

proportions of younger students are estimated to achieve better in numeracy (or literacy) than students who move into schools with high proportions of older students.

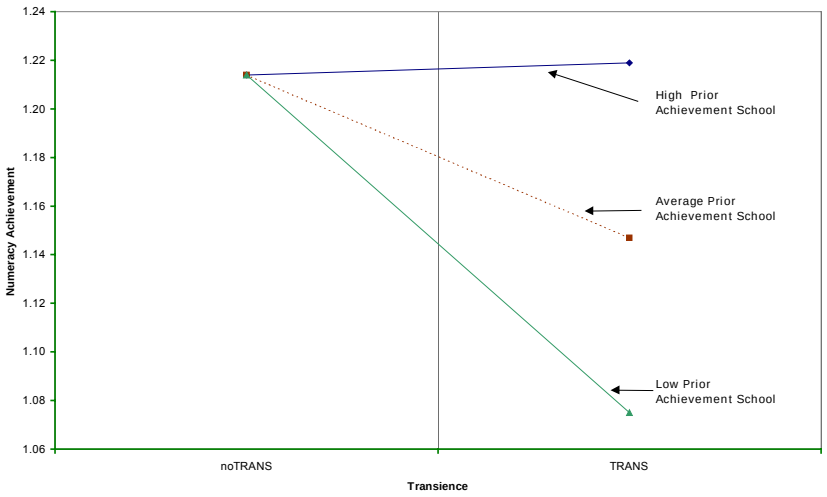


Figure 7.11 Impact of the interaction effect of student's Transience with Prior Achievement in schools on Numeracy achievement

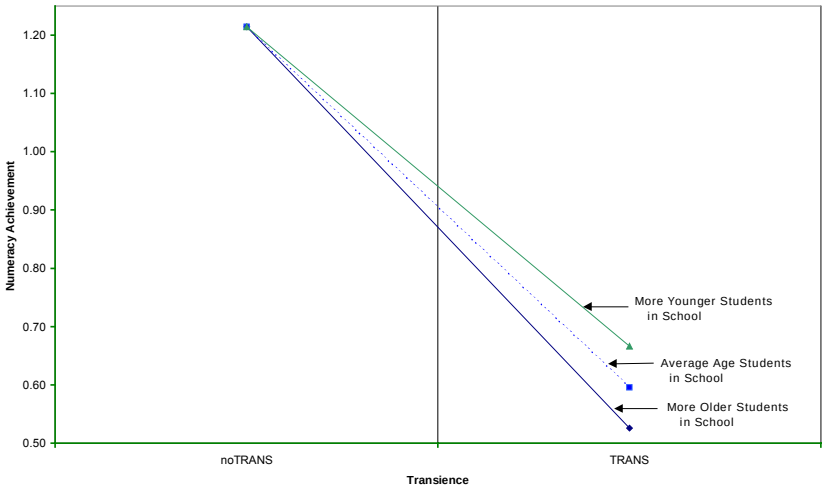


Figure 7.12 Impact of the interaction effect of student's Transience with Average Age of the Students in the school on Numeracy achievement

7. **Prior Achievement with Absenteeism Rate in Schools**: The graphical representation in Figure 7.13 shows that the difference in numeracy achievement between students with high prior achievement scores in schools with high absenteeism rates and their counterparts in schools with low absenteeism rate is

significantly ($p<0.05$) larger than the corresponding difference between students with low achievement scores. In other words, absenteeism in schools is estimated to affect achievement in numeracy of the students with high prior achievement scores more than it affects the achievement of students with low prior achievement scores.

8. **Prior Achievement with Proportion of Girls in the School**: The graph in Figure 7.14 shows that, in general, students who have high prior achievement scores in numeracy are estimated to achieve better in numeracy than students who have low prior achievement scores regardless of the proportion of girls or the proportion of boys in the school. Nonetheless, students who have high prior achievement scores in numeracy are estimated to achieve better in numeracy if they are in schools with high proportions of boys than if they are in schools with high proportions of girls. On the other hand, students who have low prior achievement scores in numeracy are estimated to achieve better in numeracy if they are in schools with high proportions of girls than if they are in schools with high proportions of boys.
9. **Prior Achievement with Average Duration of Living in Australia**: From Figure 7.15 it can be seen that regardless of the average duration of living in Australia of the students in the school, students who have high prior achievement scores in numeracy are estimated to achieve better in numeracy than students who have low prior achievement scores. However, students who have high prior achievement scores in numeracy are estimated to achieve better in numeracy if they are in schools with high proportions of students born in Australia than if they are in schools with high proportions of new students to Australia. Conversely, students who have low prior achievement scores in numeracy are estimated to achieve better in numeracy if they are in schools with high proportions of students who are new to Australia than if they are in schools with high proportions of students born in Australia.

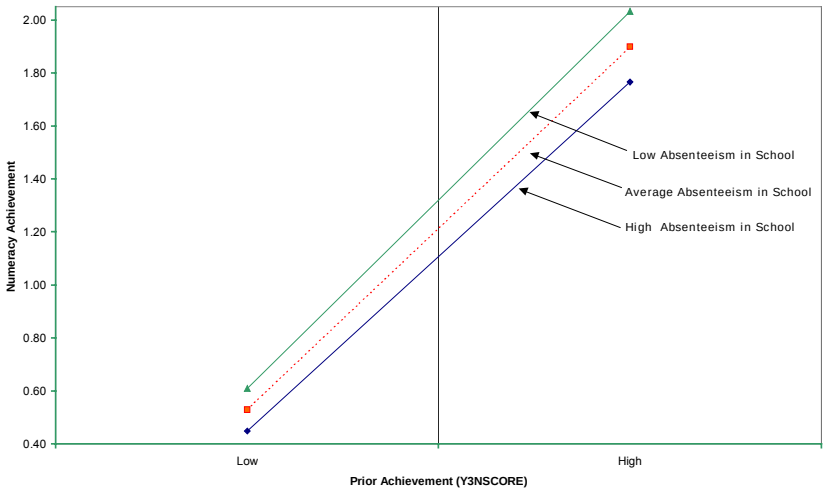


Figure 7.13 Impact of the interaction effect of student's Prior Achievement with Absenteeism Rate on Numeracy achievement

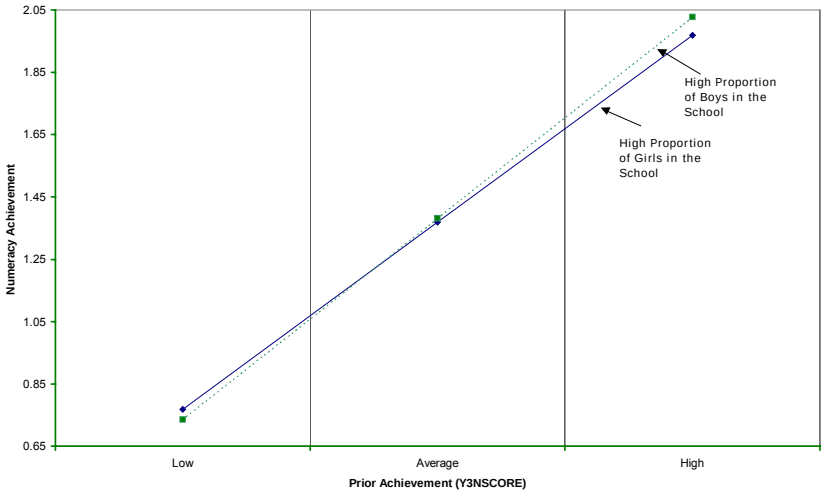


Figure 7.14 Impact of the interaction effect of student's Prior Achievement with the Proportion of Girls in schools on Numeracy achievement

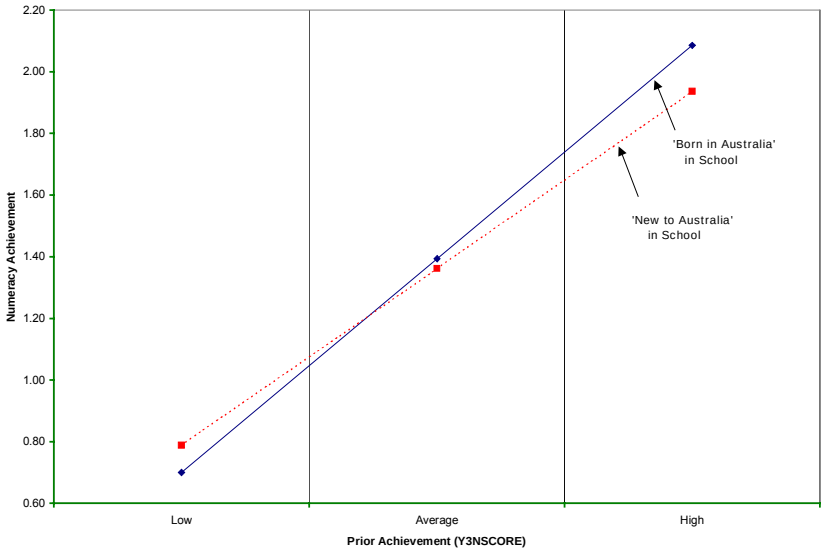


Figure 7.15 Impact of the interaction effect of student's Prior Achievement with Living in Australia in schools on Numeracy achievement

Estimation of variance explained

The results of the final estimations of the variance components for the final two-level models and the results of the variance components obtained from the null models (provided in Table 7.1) are presented together in Table 7.10 (in rows 'a' and 'b') for ease of comparison. From the information presented in rows ‘a’ and ‘b’, the information presented in rows ‘c’ to ‘f’ are calculated.

Table 7.10 Estimation of variance explained for numeracy and literacy - Two-level models

	Model-X [a]			**Model-Y** [b]			**Model-Z** [c]		
	Level-1	Level-2	Total	Level-1	Level-2	Total	Level-1	Level-2	Total
	(N=144,346)	(N=2,868)		(N=37,832)	(N=1,853)		(N=32,741)	(N=1,823)	
Numeracy									
a) Var. Comp. Null Model	1.17	0.39	1.56	1.00	0.22	1.22	0.99	0.21	1.20
b) Var. Comp. Final Model	1.14	0.32		0.56	0.06		0.54	0.06	
c) Var. Available	75.1%	24.9%		81.7%	18.3%		82.1%	17.9%	
d) Var. Explained	2.9%	17.1%		44.3%	72.1%		45.2%	70.7%	
e) Total Var. Explained	2.2%	4.3%	6.4%	36.2%	13.2%	49.4%	37.1%	12.6%	49.7%
f) Var. Left Unexplained	72.9%	**20.6%**	93.6%	45.5%	**5.1%**	50.6%	45.0%	**5.2%**	50.3%
Literacy									
a) Var. Comp. Null Model	1.21	0.32	1.53	1.00	0.22	1.22	0.97	0.22	1.19
b) Var. Comp. Final Model	1.16	0.26		0.45	0.05		0.44	0.05	
c) Var. Available	79.0%	21.0%		81.8%	18.2%		81.7%	18.3%	
d) Var. Explained	3.5%	17.7%		54.6%	75.6%		54.7%	75.1%	
e) Total Var. Explained	2.7%	3.7%	6.5%	44.6%	13.8%	58.4%	44.7%	13.8%	58.5%
f) Var. Left Unexplained	76.3%	**17.3%**	93.5%	37.1%	**4.4%**	41.6%	37.0%	**4.6%**	41.5%

Note: a - For Model-X, the simplest model has the variable YEARLEVL as the only predictor.
b - Transience model
c - Non-transience model

A discussion of the calculations involved in Table 7.10 is presented in Chapter 4 and can also be found in Raudenbush and Bryk (2002; pp.68-95) and in Bryk and Raudenbush (1992; pp.60-76).

For example, for literacy, the predictors included in the final transience model (Model-Y), explain 54.6 per cent of the 81.8 per cent variance available at the student-level, and that is equal to 44.6 per cent (that is, 54.6×81.8) of the total variance explained at the student-level. Similarly, for the same model, the predictors included in the final model explain 13.8 per cent (that is, 75.6 per cent of 18.2 per cent) at the school-level. Therefore, the total variance explained by the predictors included in the final two-level transience model for literacy is $44.6 + 13.8 = 58.4$, which leaves 41.6 per cent of the total variance unexplained. Thus, for numeracy and literacy, the results in Table 7.10 (row 'f') show that the percentages of variances left unexplained at the student-level are much larger compared with the percentages of variances that are left unexplained at the school-level, regardless of the type of model tested. However, for both subjects, the percentages of total variances that are left unexplained (in the shaded cell of row 'f') in Model-X are much larger compared with the percentages of total variances that are left unexplained in either Model-Y or Model-Z. In addition, the percentages of total variances that are left unexplained in Model-Y follow closely the percentages of total variances that are left unexplained in Model-Z for numeracy as well as literacy.

In general, the percentages of total variances explained in the final Models Y and Z have not increased considerably compared with the percentages of variances that were explained by Prior Achievement only (see results in Table 7.4). Obviously, the inclusion of other predictors at the student-level and at the school-level has done relatively little to increase the amounts of variances explained in these final models.

It is worth noting that the percentages of variances that are explained in the final Models Y and Z are noticeably large especially at the school-level where about 70 per cent or more of the available variances are explained. Consequently, around five per cent of the total variances that are available at the school-level are left unexplained (given in bold row 'f'). Furthermore, the percentages of variances that were explained at the school-level in these two types of models by Prior Achievement alone (see results in Table 7.4) are almost equal to the percentages of total variances that are explained in the final models (results in Table 7.10). Hence, it seems that the inclusion of Prior Achievement alone is enough to reduce substantially the amount of variance available for explanation at the school-level in Models Y and Z.

Finally, for Models Y and Z, it is also worth noting that the total variances that are left unexplained (in the shaded cell of row 'f') for numeracy are noticeably larger compared to the total amounts of variances that are left unexplained for literacy. For example, for Model-Y, the percentage of total variance that is left unexplained for numeracy is 50.6 while the corresponding percentage for literacy is 41.6. The same observation is also evident at the school-level, where slightly more variances are left unexplained for numeracy than for literacy. Thus, it appears that most of the important school-level factors influencing achievement in literacy of the Grade 5 students in South Australia have been included in the models developed here.

Comparison of model fit using the deviance statistic

It was said in Chapter 4 that the deviance statistic and a chi-square test could be used to compare the fit of a series of models that are subsets of a more complex model. It

was also noted that this chi-square test is best used if the Full Maximum Likelihood (MLF) procedure is employed as the estimation mode, and not when the Restricted Maximum Likelihood (MLR) estimation procedure is used (Raudenbush et al., 2000). This is because under the MLR estimation procedure, the number of parameters remains the same between two models which differ only in their regression coefficients and therefore the chi-square test can only be used to determine the fit of the unconditional part of the model.

In the analyses described above, the MLR procedure was employed because Raudenbush and Bryk (2002; p.53) have pointed out that "the MLR estimates of variance components do adjust for uncertainty about the fixed effects, and the MLF results do not". Nevertheless, the selection of the better estimation procedure to employ at this stage was not considered critical because Raudenbush and Bryk (2002; p.53) have further indicated that for "two-level models, MLF and MLR will generally produce very similar results for δ^2, but noticeable differences can occur in estimation of T". Besides, they have argued that in cases where the number of Level-2 units is large (as is the case with the number of schools here), the two procedures will produce very similar results.

However, for purposes of comparing the fit of the models using the deviance statistics, the same HLM analyses described above are repeated this time using the MLF procedure. For each of the HLM runs, the chi-square test described above is used to compare the fit of a model with the preceding model. At this stage, an optional hypothesis testing sub-routine available in HLM5/2L is employed to compare model fit in successive HLM runs. This is done by entering the deviance statistic and number of parameters reported in the output file of a previous model into the optional hypothesis testing dialog box fields provided in HLM5/2L. A chi-square statistic, with associated degrees of freedom and p-value are then estimated and printed at the end of the next HLM5/2L output file.

It should be noted that for each of the predictors in the final models (Figures 7.1 to 7.6), its inclusion in the model results in significant improvement of the model fit as indicated by the chi-square test printed in the output generated using the MLF procedure. Furthermore, the results (that is, fixed effects and t-ratio values as well as the estimates of variance components) obtained using the MLF and MLR procedures are very similar and in most cases differ only in the second or the third digit after the decimal point.

Table 7.11 present results of deviance statistics and chi-square tests carried out to compare model fit at the conclusion of the unconditional part and at the conclusion of the final models for numeracy and literacy for the three types of models examined. In Table 7.11, the fit of the unconditional model is compared to the fit of the null model (or grade-level-only model for Model-X), and the fit of the final model is compared to the fit of the unconditional model.

The first two columns of Table 7.11 compare the fit of the models using the deviance statistics obtained with the MLR procedure while the other five columns compare the fit of the same models using the chi-square tests obtained with the MLF procedure.

For example, the results in Table 7.11 indicate that the value of the deviance statistic obtained using the MLR procedure from the null model for numeracy in Model-Y is 110,113.68, and that obtained from the Level-1 model is 88,314.33 resulting in a drop of deviance of 21799.34. Similarly, there is a drop of 339.25 in the size of the deviance obtained from the final model compared to the value obtained from the unconditional model. For the same model but using the MLF procedure Table 7.11 indicates that the value of deviance statistic obtained from the null model is

110,108.66 with three estimated parameters, and that obtained from the Level-1 model is 88,265.59 with 36 estimated parameters. Hence, the change in deviance (as indicated by the chi-square value) is 21843.07 with 33 degrees of freedom and an associated p-value of 0.00, which shows improved fit of the model. Likewise, there is improved fit of the final model compared to the Level-1 model as indicated by the chi-square value of 393.61 with ten degrees of freedom and a significant p-value. The results in Table 7.11 show similar findings for all the other models examined for numeracy and literacy. Hence, the inclusion of the student-level predictors and school-level predictors significantly improve the overall fit of the models. [See also Hungi (2003; pp.219-220, pp.497-499 and p.529) for a discussion on examination of residuals and the adequacy of the log transformation].

Table 7.11 Comparison of model fit using the chi-square tests

a) Model-X

	Using MLR		**Using MLF**				
	Deviance Statistic	Change in Deviance	Deviance Statistic	Number of Parameters	Chi-square Statistic	Degrees of Freedom	p-value
Numeracy							
Grade-level-only	441,640.29		441,624.31	6			
Unconditional	437,117.95	4,522.34	437,080.76	21	4,543.55	15	0.00
Final	435,804.04	1,313.91	435,705.38	26	1,375.38	5	0.00
Literacy							
Null	448,361.86		445,000.69	6			
Unconditional	439,536.94	8,824.92	439,498.98	21	5,501.71	15	0.00
Final	437,909.38	1,627.56	437,794.37	28	1,704.60	7	0.00

b) Model-Y

	Using MLR		Using MLF				
	Deviance Statistic	Change in Deviance	Deviance Statistic	Number of Parameters	Chi-square Statistic	Degrees of Freedom	p-value
Numeracy							
Null	110,113.68		110,108.66	3			
Unconditional	88,314.33	21799.34	88,265.59	36	2,1843.07	33	0.00
Final	87,975.09	339.25	87,871.98	26	393.61	10	0.00
Literacy							
Null	110,069.26		110,064.24	3			
Unconditional	80,247.65	29821.61	80,188.01	45	29,876.23	42	0.00
Final	80,112.21	135.44	80,002.57	23	185.44	22	0.00

c) Model-Z

	Using MLR		Using MLF				
	Deviance Statistic	Change in Deviance	Deviance Statistic	Number of Parameters	Chi-square Statistic	Degrees of Freedom	p-value
Numeracy							
Null	94,978.15		94,973.15	3			
Unconditional	75,546.48	19431.68	75,546.48	22	19,426.68	19	0.00
Final	75,349.45	197.03	75,257.81	18	288.67	4	0.00
Literacy							
Null	94,598.94		94,593.95	2			
Unconditional	68,302.10	26296.84	68,252.08	36	26,341.86	34	0.00
Final	68,199.13	102.97	68,103.68	18	148.40	18	0.00

Discussion of factors influencing student achievement

With only a few exceptions, the results reported in this chapter are generally consistent with what has been found in past studies in Australia and overseas regarding student-level and school-level factors that have significant influences on student achievement. This point is expounded in the following paragraphs, which give examples of past studies that looked at the factors found to have significant influences on student achievement in this chapter.

For Age of Student, Afrassa and Keeves (1999) using South Australia BSTP data for 1995, 1996 and 1997 found that younger students within a grade were likely to perform better on the BST than their older counterparts, which is consistent with the results reported in this chapter. Contrary to what has been found in the current study, Peck and Trimmer (1995) reported that teachers in Western Australia contended that younger students tended to thrive less well than older members of the class. In addition, Peck and Trimmer reported that younger students were more likely than older students to repeat a class before reaching Grade 12, and that the younger students who repeated a class achieved much less than those who made normal progress. Likewise, another study in Australia by Griffin and Harvey (1995) found that younger children had more problems academically and socially and tended to remain behind their older classmates in achievement. Nevertheless, Peck and Trimmer (1995) found that the differences in achievement between younger and older students tended to be reduced in subsequent years and that at Grade 12 the younger students were likely to achieve as well as older students in the same class who started school at the same time.

For Sex of Student, consistent with the results reported in this chapter, many studies have reported gender differences in achievement. In summarizing the research findings from 35 years of IEA studies regarding sex differences, Keeves (1995; p.23) concluded "differences are found between sexes in achievement which vary in size and direction across countries, school and subjects and overtime". He adds that gender differences would appear to be related to societal and curricular factors and not to genetic factors, as has sometimes been assumed. Consistent with what is reported in this chapter, in many studies boys are reported to be outperforming girls in mathematics (e.g. Husén, 1967; Bishop and Clement, 1994; Goh and Fraser 1996) while girls are reported to be outperforming boys in literacy (e.g. Thorndike, 1973b; Braggett, 1997; Yeung and Marsh, 1997; Lokan et al., 2001).

Evidence of occurrence of gender differences in achievement in South Australia primary schools is available. Teachers in South Australia government primary schools in 1997 were asked to assign their Grades 1 to 8 students to levels of achievement using the nationally developed curriculum profiles in English, Science, Studies of Society and Environment (SOSE), and Technology. Rothman (1998) reported that girls achieved higher levels in every strand (reading, viewing and listening) of English and at every grade level than boys in the same grade. Similar observations were made for SOSE, although in this case there was little gender difference in achievement for the lower primary Grades 1 to 4 students. On the other hand, boys achieved at a higher level in science and technology than girls, except at Grade 3 level where there were little differences between boys and girls. In addition, at the Grades 3 and 5 levels in South Australia, the study by Afrassa and Keeves (1999) reported significant gender differences in numeracy and literacy achievement. However, Afrassa and Keeves (1999) reported that the role of gender in student achievement was unclear because

they found that the sign of regression coefficient for the gender variable was not consistent in the models that they examined.

For Racial Background (ATSI), Australian studies that have investigated the effects of Aboriginality on academic achievement have reported findings that are consistent with what is reported in this chapter. There are clear indications that, in Australia at the primary school level, the achievements in numeracy and literacy of ATSI students are much lower than that of the general population. Indeed, reports from major Australian studies note that Aboriginality as an important predictor of achievement (e.g. ASSP[19] [Keeves and Bourke, 1976], VQSP[20] [Hill, 1996], NSELS[21] [Masters and Forster, 1997], TIMSS[22] [Lokan et al., 1997)] RWAGSS[23] and WASES[24] [Young, 1998a]). For example, Masters and Forster (1997) reported that the average literacy levels of Grades 3 and 5 ATSI students were three to four years below the average of students in the main sample in the National School English Literacy Survey. Lokan et al. (1997) reported similar differences for numeracy at the primary school level. In South Australia, Afrassa and Keeves (1999) reported similar findings in both numeracy and literacy at the Grade 3 level as well as at the Grade 5 level. Studies at the secondary school level in Australia have also documented the lower average literacy and numeracy levels of ATSI students (Marks and Ainley, 1997; Lokan et al., 1996 & 1997; Lokan and Greenwood, 2001).

For English speaking background, past studies have reported inconsistent relationships between NESB and achievement in school learning in Australia. Martin and Meade (1979) reported on a comparative study of migrant students of NESB origin and students whose parents were born in an English speaking country. The study revealed that a substantially greater proportion of children of NESB origin than of Australian or other English speaking origin achieved creditable Higher Schools Certificate (HSC) results. The study further revealed that, when similar I.Q. and socioeconomic status groups were compared, students of non-English speaking migrant background did as well or better than other students in terms of School Certificate results.

On the other hand, Keeves and Bourke (1976) reported higher levels of performance in numeracy and literacy among 10-year-old students from English (or Northern European language) background compared to students of other language backgrounds, which is consistent with what is found in this chapter. In addition, Ainley et al. (1990) reported that reading comprehension in primary schools in Victoria was related to the English speaking background of the student, with students who had both parents from a non-English speaking background, or who themselves were born in non-English speaking country having lower achievement. The findings were similar for mathematics though not as strong. Similarly, results of the basic skills tests in literacy and numeracy administered to the Grades 3 and 6 students in primary schools in New South Wales in 1989, showed a clearly poorer performance of children from NESB that could not be attributed to the tests' cultural bias (Davies, 1991). Moreover, recent studies have documented this same relationship between NESB status and performance levels among students at Grades 3 and 5 (Masters and Forster, 1997) and among secondary school students (Lokan et al., 2001; Marks and Ainley, 1997). In contrast, PISA[25], TIMSS and TIMSS-R 1999 data suggest that NESB status has no

[19] Australian Studies in School Performance.
[20] Victoria Quality School Project.
[21] National School English Literacy Survey.
[22] Third International Mathematics and Science Study.
[23] Rural Western Australia Government Schools Survey.
[24] Western Australian School Effectiveness Study.
[25] Programme for International Student Assessment.

significant influence on mathematics achievement in middle primary and junior secondary schools in Australia (Mullis et al., 1997 & 2000; Lokan et al., 2001).

For Living in Australia or migrant status (INOZ), the OECD[26] PISA study, which measured levels of performance in numeracy and literacy among 15-year-old in 32 countries, found that in most countries migrant students have lower average levels of achievement when compared to so-called 'native' students. However, this was not the case in Australia, where the migrant status of the student did not influence the performance levels in numeracy and literacy (OECD, 2001). For South Australia at the primary school level, Afrassa and Keeves (1999) reported that, when other variables were held equal, migrant students outperformed students born in Australia in numeracy (but not necessarily in literacy) in some of the models they examined. Nevertheless, when interpreting the effects of migrant status on student achievement in Australia it should be borne in mind that migration to Australia is based on skills. As a consequence, a vast majority of new students to Australia are from more highly educated parents and arguably from high socioeconomic status home backgrounds.

For Absenteeism, although this variable is only available for examination at the group-level in the analyses reported in this chapter, it is generally argued that there is a strong correlation between attendance and academic success. Obviously, students who are regular absentees receive fewer hours of instruction and therefore are highly likely to achieve less compared to the rest of their classmates. In the South Australia context at the primary school level, Rothman (2000, 2001 & 2002) argues that high rates of absenteeism affected regular attendees as well, because teachers must accommodate non-attendees in the same class. Thus, the results of the analyses reported in this chapter would appear to confirm Rothman's argument.

For School-card (socioeconomic status), studies in Australia and overseas agree that socioeconomic status has a significant influence on student achievement, with students from wealthy homes doing better than students from impoverished homes. In South Australia, Rothman (1998) reported that in 1997 non-school cardholders were found to have achieved at higher levels than school cardholders in most areas of learning (except English at Grade 3 and below) in government primary schools in South Australia. In addition, Afrassa and Keeves (1999) found that South Australian Grades 3 and 5 students in schools with higher proportions of school cardholders achieve less well in both numeracy and literacy than their counterparts in school with lower proportions of school cardholders.

For Transience or Mobility, there are very few studies that have examined the influence of this factor on academic achievement at the primary school level in Australia. However, overseas research evidence indicates that this factor has a negative effect on student progress in school (e.g. Brent and Diobilda, 1993; Rumberger and Larson, 1998; Wright, 1999; Temple and Reynolds, 1999; Reynolds and Wolfe, 1999), which is consistent with the results reported in this chapter. In Australia, Fields (1995) found mobile students experience both academic and social difficulties. In addition, Hill (1996) reported that a major study (School Global Budget Research Project) identified transience as a powerful predictor of school learning in Australia with negative effects.

For School Location (urban or rural), there are few studies that have been conducted in Australia to investigate the influence of this factor on student achievement at the primary school level. Nevertheless, the available results pertaining to the relationship between school location and student achievement indicate that there are only small or

[26] Organisation for Economic Co-operation and Development.

negligible differences in the average numeracy and literacy levels of students attending primary schools in different areas in Australia, and likewise secondary schools. For example, for primary school 10-year-old students, the Australian Studies of Student Performance (ASSP) found that metropolitan students displayed marginally higher levels of numeracy and literacy than non-metropolitan students in 1975 (Keeves and Bourke, 1976) but not in 1980 (Bourke et al, 1981). More recently, the NSELS study found that students in Grades 3 and 5 in major urban areas had higher levels of literacy achievement than their counterparts in small rural centres, but the differences were not large (Masters and Forster, 1997). Similarly, in South Australia, the study by Afrassa and Keeves (1999) found that the locality of the school had a small (though significant) effect on performance in literacy, but not numeracy at the Grades 3 and 5 levels, with urban students performing at a higher level than their rural counterparts. These results are generally consistent with the results reported in this chapter.

At the secondary school level, the ASSP studies found very small differences in the average numeracy and literacy levels favouring metropolitan students among 14 year olds in both 1975 and 1980 (Keeves and Bourke, 1976; Bourke et al., 1981). Similarly, results from the PISA, and TIMSS studies pertaining to the Australia data show that there are few differences in mathematics and science achievement of students attending schools in different areas (Lokan et al., 2001; Webster and Fisher, 2000). The Longitudinal Surveys of Australian Youth (LSAY) project has also reported negligible differences in numeracy and literacy achievements among 15-year students in metropolitan schools and their counterparts in rural schools (Hillman, Marks and McKenzie, 2002).

However, an earlier study by Young (1998b) demonstrated that the location of the secondary school had a significant effect upon student achievement, with students attending rural schools not performing as well as students from urban schools in the areas of Mathematics and Sciences in Western Australia.

For School Size, considerable disagreement exists in the literature relating to the effects of this factor on student achievement. Some studies claim that students perform better in small schools (Jolly and Deloney, 1993; Raywid, 1997), some find no difference (Ramilez, 1990; Plecki, 1991; Luyten, 1994b; Lamdin, 1995), whereas others claim that students in larger schools perform better (McKenzie, 1988; Mok and Flynn, 1996; Bourke, 1998).

Mok and Flynn (1996) reported that Grade 12 students from larger Catholic schools in New South Wales, on average tended to achieve at a higher level than their counterparts from smaller schools, even after controlling for the students' background, motivation and school culture variables using multilevel analysis techniques. However, McKenzie (1988) argued that although big schools bring considerable advantages to student achievement, plateau effects begin to appear at relatively low enrolment levels. At the primary school level in South Australia, the study by Afrassa and Keeves (1999) found that school size did not have a significant influence on student achievement in numeracy and literacy in any of the models that they examined, which is not consistent with the results reported in this chapter. However, it should be mentioned that the models examined in this chapter differ from those examined by Afrassa and Keeves. In the study by Afrassa and Keeves, BSTP data from each of the three testing occasions of interest in that study (1995, 1996 and 1997) were analyzed separately and Grade 3 data were analyzed separately from Grade 5 data. In this chapter, BSTP data from six testing occasion of interest in this study (1995 to 2000) for both grade levels are analyzed simultaneously.

Conclusions

The results of Model-X two-level HLM analyses for both numeracy and literacy indicate that at the student-level, there are four variables that have significant influences on achievement in numeracy and literacy out of the six variables in the proposed models. The same four student-level variables that have a significant influence on achievement in numeracy also have a significant influence on achievement in literacy. The four variables are (a) Grade Level (YEARLEVL), (b) Age of the Student (AGE), (c) Racial Background (ATSI), and (d) Speaking English at Home (HOME).

On the other hand, the results of Model-Y two-level HLM analyses indicate that there are six student-level variables that have a significant influence on achievement in both numeracy and literacy, namely SEX (Sex of the Student), AGE, ATSI, INOZ (Living in Australia), TRANS (Transience) and Y3NSCORE (or Y3LSCORE for literacy). Apart from TRANS, these are the same student-level variables that have a significant influence on achievement in numeracy and literacy in Model-Z. In addition, Models Y and Z results indicate that the variable Speaking English at Home (HOME) has a significant influence on achievement in literacy but not in numeracy at the student-level.

By and large, the results at the student-level for the three types of models generally agree regarding which of student-level variables examined in this study have a significant influence on achievement in the BST among Grades 3 and 5 students in South Australia when other variables are equal.

At the school-level, the results of analyses of three types of models also generally agree regarding which of the school-level variables examined in this study have direct significant influences on the achievement in numeracy and literacy of the Grades 3 and 5 students in South Australia. For instance, the results indicate that there are four variables that have direct significant influences on student achievement in both numeracy and literacy in all the proposed models. The four variables are (a) Proportion of School-card Holders (PSCARD), (b) School Location (either METRO or GPOLOG), (c) Mobility Rate (either MOBILITY or TRANS_1), and (d) Absenteeism Rate (ABSENT). In addition, the Model-X results indicate that the School Size variable (SSIZELOG) has a significant influence on achievement in both numeracy and literacy. Then again, Model-Y results indicate that this School Size variable as well as the variable ATSI_1 (Proportion of non-ATSI Students) have significant influences on achievement in numeracy but not in literacy.

The results also indicate that of the 12 student-related variables examined at the school-level, seven had no direct influence on achievement in numeracy and literacy in any of the models examined. The seven variables are SEX_1 (Proportion of Girls), AGE_1 (Average Age), HOME_1 (Average Speaking English at Home), INOZ_1 (Average Living in Australia), CAP (Country Area Program), YR35PPT (Participation Size) and Y3NSCO_1 (Prior Achievement for numeracy) or Y3LSCO_1 (Prior Achievement for literacy). Four of these seven school-level variables (SEX_1, AGE_1, INOZ_1 and Y3NSCO_1 or Y3LSCO_1) have interaction effects with the student-level variances and therefore they have indirect influence on achievement in numeracy and literacy. However, three of the seven variables (namely CAP, YR35PPT and HOME_1) have no interaction effects with any of the student-level variables in any of the models examined.

For Models Y and Z, the results also indicate that the total variance to be explained for numeracy and literacy at the student-level are much larger (around 82 per cent) compared to the total variance to be explained at the school-level (around 18 per

cent). Results similar to these were also obtained based on Model-X. However, in the final models, the amounts of variance that are explained in Models Y and Z are much larger compared to the amounts of variance that are explained in Model-X. Furthermore, in Models Y and Z, the amounts of total variance explained by Prior Achievement alone are almost equal to the total amount explained in the final models. Moreover, in these two types of models, the total amounts of variance that are explained by Prior Achievement alone at the school-level are almost equal to the amounts of variance explained in final models. These results indicate that the Grade 3 scores can explain much of the variance that is available in the Grade 5 scores at the school-level.

If the total amounts of variance explained in the final model were to be used as a measure of how good a model is in representing the relationship between the factors involved in student performance, then Models Y and Z are the better models compared to Model-X. Using this criterion, it is not possible to separate Model-Y from Model-Z because the total amounts of variance explained in the two models are basically the same.

Finally, the results and discussion presented in this chapter have indicated that some of the dummy variables used to specify the different testing occasions (or cohorts) and the linear trend variable (OCC) have influences on achievement in numeracy and literacy in some of the models examined. These results may be interpreted to mean that students' achievements in numeracy and literacy are influenced by the testing occasion, that is, there are advantages or disadvantages associated with taking the tests on a particular occasion. However, these results could also be interpreted to mean that students on a particular testing occasion had significantly better (or poorer) mastery of the basic skills of numeracy and literacy than on the other occasions. Furthermore, these results could also be indicating the existence of possible equating errors embedded in the scores. Nevertheless, it is not possible to distinguish between the situations where there was markedly better or poorer mastery of the skills from situations where the scores could have been distorted in the equating process.

Regardless of the situations outlined above that might have applied, it should not be forgotten that in the two-level HLM analyses reported in this chapter, schools are not linked over time. Each school is treated as a different entity on each testing occasion, and therefore the advantages or disadvantage mentioned above may not necessarily be noticeable at the school-level. Obviously, using the two-level analyses it is not possible to obtain a clear picture of what is happening to the performance at the school-level over time, and therefore, there is the need for further analyses.

The next chapter re-examines the three models using the HLM5/3L computer program. For the current study, the main advantage of HLM5/3L over HLM5/2L is that the former computer program allows the identity of the school to be kept intact over time. In addition, HLM5/3L disentangles the amounts of variance available and explained at the occasion-level from the amounts of variance available and explained at the school-level, and might give a better representation of the whole system.

8

Achievement Factors: Three-level Models

In the two-level analyses reported in the previous chapter, unique identities were employed for each Level-2 unit included in the analyses. This meant that each school was treated as a different school on each testing occasion. Because of the multilevel nature employed in the two-level analyses, at the planning stages of this study there were some concerns about the appropriateness of employing the two-level models described in the previous chapter to study the factors influencing student achievement. In particular, there were concerns about the appropriateness of partitioning of variance and monitoring of linear (or quadratic) trends in achievement using the two-level models given that the multilevel nature employed in those models did not link data of the same school from the different testing occasions. For that reason, a decision was made to reformulate the three models (Models X, Y and Z) in terms of a three-level structure and, accordingly, employ the HLM5/3L (Raudenbush et al., 2000) computer program.

Thus, this chapter presents the examination of the same three models (Models X, Y and Z) using the HLM5/3L (Raudenbush et al., 2000) computer program. The multilevel structure employed in this program allows the identity of the school to be kept intact over time.

The general framework of this chapter is similar to that of the previous chapter. That is, a description of the models to be examined is provided and then the analyses using the HLM5/3L computer program are described. However, most of the details involved in the HLM analyses are not provided in this chapter because these details have already been introduced in the previous chapters. In addition, sections on the effects of Grade Level and Prior Achievement, and the section on cross-level interaction effects that were included in the previous chapter have been omitted in this chapter because they were lengthy and basically provided no additional information.

The concluding section in this chapter compares the results of the two-level analyses reported in the previous chapter with the results of the three-level analyses reported in

this chapter. However, where relevant, such comparisons are also provided in the main text of the chapter.

Descriptions of the three-level HLM models

The three three-level hierarchical linear models proposed for teasing out factors influencing student achievement in this study are introduced in Chapter 5 (Figures 5.5, 5.6 and 5.7 respectively). The data sets and the variables examined for inclusion in each of the three models are also described in Chapter 5 and therefore it is not necessary to repeat them here.

It should be remembered that the three three-level models examined in this chapter correspond directly to the three two-level models (Models X, Y and Z) that are examined in the previous chapter. Thus, Model-X uses all the students who have taken part in the BSTP from 1995 to 2000, while Model-Y (transience) uses all those students who could be matched, and Model-Z (non-transience) use only those students who could be matched and who remained in the same schools over the two-year period. At the student-level, the structures of the three-level models are exactly the same as the structures of the two-level models. However, at the school-level, the three-level models do not include the dummy variables denoting the testing occasions or student cohorts (OCC1 to OCC6) and the trend variables (OCC and OCCSQD), and instead these variables are included in a level of their own, the occasion-level or macro-level, a third level.

The steps undertaken in the three-level HLM analyses are similar to those undertaken in the two-level HLM analyses. The first step is to run the null models to obtain the estimates of the amounts of variances to be explained in the models, that is without entering into the equations any student-level, school-level and occasion-level variables. The second step is to run the unconditional model at the micro-level, that is adding into the equation the significant student-level variables, but without any school-level and occasion-level predictors. The third step is to run the school-level or meso-level model, that is entering into the equations the significant student-level variables and the significant school-level variables, together with the significant variables for interaction effects, but without any occasion-level variables. And the fourth step is to run the final model by entering into the macro-level model the significant occasion-level variables.

Specifications of the three-level null models

For the current study, and following the procedure as well as the symbols given by Bryk and Raudenbush (1992; p.176), the simplest three-level models to represent how variation in the outcome variables is allocated across the three different levels (student, school, and occasion), can be described as follows:

Level-1 model

At the student-level, the student achievement is modelled as a function of a school mean plus a random error:

$$\mathbf{Y}_{ijk} = \boldsymbol{\pi}_{0jk} + \mathbf{e}_{ijk} \qquad \text{Equation 8.1}$$

where:

$\mathbf{Y}_{ijk}$ is the achievement (Rasch score) of student i in school j and occasion k;

$\boldsymbol{\pi}_{0jk}$ is the mean achievement of school j on occasion k; and

$\mathbf{e}_{ijk}$ is a random error or 'student effect', that is, the deviation of the student mean from the school mean.

The indices *i*, *j* and *k* denote students, schools and occasions where there are:

$i = 1, 2, \ldots, n_{jk}$ students within school j on occasion k;

$j = 1, 2, \ldots, J_k$ schools within occasion k; and

$k = 1, 2, \ldots, K$ occasions.

A simplified form of Equation 8.1 is presented in the output file generated by HLM5/3L computer program, where **Y**, **P0** and **E** are used to represent the components $\mathbf{Y}_{ijk}$, $\boldsymbol{\pi}_{0jk}$ and $\mathbf{e}_{ijk}$ in Equation 8.1, respectively. Hence, the Level-1 null model equation in the output file becomes:

$$\mathbf{Y} = \mathbf{P0} + \mathbf{E} \qquad \text{Equation 8.2}$$

Level-2 Model

At the school-level, each school mean, $\boldsymbol{\pi}_{0jk}$, is viewed as an outcome varying randomly around some occasion mean.

$$\boldsymbol{\pi}_{0jk} = \boldsymbol{\beta}_{00k} + \mathbf{r}_{0jk} \qquad \text{Equation 8.3}$$

where:

$\boldsymbol{\pi}_{0jk}$ is the mean achievement of school *j* on occasion *k*; and

$\boldsymbol{\beta}_{00k}$ is the mean achievement on occasion *k*,

$\mathbf{r}_{0jk}$ is a random 'school effect', that is, the deviation of a school mean from the occasion mean. Within each of the occasions, the variability among schools is assumed to be the same.

A simplified form of Equation 8.3 that is presented in the output file generated by HLM5/3L computer program is:

$$\mathbf{P0} = \mathbf{B00} + \mathbf{R0} \qquad \text{Equation 8.4}$$

where:

P0, **B00** and **R0** are used to represent the components $\boldsymbol{\pi}_{0jk}$, $\boldsymbol{\beta}_{00k}$ and $\mathbf{r}_{0jk}$ in Equation 8.3, respectively.

Level-3 Model

At the occasion-level, each occasion mean $\boldsymbol{\beta}_{00k}$ is viewed as varying randomly around a grand mean:

$$\boldsymbol{\beta}_{00k} = \boldsymbol{\gamma}_{000} + \mathbf{u}_{00k} \qquad \text{Equation 8.5}$$

where:

$\boldsymbol{\beta}_{00k}$ is the mean achievement on occasion *k*,

$\boldsymbol{\gamma}_{000}$ is the grand mean

$\mathbf{u}_{00k}$ is a random 'occasion' effect, that is, the deviation of occasion mean from the grand mean.

A simplified form of Equation 8.5 presented in the output file generated by HLM5/3L computer program is:

$$\mathbf{B00} = \mathbf{G000} + \mathbf{U00} \qquad \text{Equation 8.6}$$

where:

B00, **G000** and **U00** are used to represent the components β_{00k}, γ_{000} and u_{00k} in Equation 8.5, respectively.

Variance partitioning

Table 8.1 displays estimates of the variance involved in the three-level models for numeracy and literacy. The percentages of variance available at each of the two levels of hierarchy are calculated from the variance components by employing the formulae presented in Chapter 4.

For Model-X, it should be noted that the simplest three-level models have the variable YEARLEVL as the only predictor and therefore the results of variance partition displayed in Table 8.1 for Models Y and Z are providing better pictures of the situations for student scores than the results for Model-X.

Table 8.1 Variance partitioning using the three-level models

Term		Model-X[a]		Model-Y[b]		Model-Z[c]	
		Variance Component	(%) Var. Available	Variance Component	(%) Var. Available	Variance Component	(%) Var. Available
Numeracy							
Student	σ_0^2	1.18	(75.0)	1.00	(81.8)	0.99	(82.2)
School	$\tau_{\pi 0}$	0.39	(24.8)	0.22	(18.1)	0.21	(17.6)
Occasion	$\tau_{\beta 0}$	0.00	(0.2)	0.00	(0.2)	0.00	(0.2)
Total =	$\sigma_0^2 + \tau_{\pi 0} + \tau_{\beta 0}$	**1.57**		**1.22**		**1.20**	
Literacy							
Student	σ_0^2	1.18	(75.0)	1.00	(81.9)	0.98	(81.8)
School	$\tau_{\pi 0}$	0.39	(24.8)	0.20	(16.3)	0.19	(16.3)
Occasion	$\tau_{\beta 0}$	0.00	(0.2)	0.02	(1.8)	0.02	(1.9)
Total =	$\sigma_0^2 + \tau_{\pi 0} + \tau_{\beta 0}$	**1.57**		**1.22**		**1.19**	

Note: a - For Model-X, the simplest model has the variable YEARLEVL as the only predictor.
b - Transience model
c - Non-transience model

Based on Model-Y (the transience model), the results in Table 8.1 show that 81.8, 18.1 and 0.2 per cent of the variation of Grade 5 pupils' numeracy scores are at the student, school and occasion levels respectively, and based on Model-Z (the non-transience model) the percentages are 82.2, 17.6 and 0.2 respectively. The corresponding percentages for students' literacy scores based on the transience model are 81.9, 16.3, and 1.8 for student, school and occasion levels respectively, and the corresponding percentages based on the non-transience model are 81.8, 16.3 and 1.9 respectively.

For all the three proposed models, the results in Table 8.1 indicate large variation between students within schools in terms of their achievement in numeracy and literacy compared to the variation in performance among schools. The results also indicate the existence of very little variation of schools in terms of their students' performance in numeracy and literacy between occasions.

Generally, for both outcome measures, the variation of student scores at the student-level as well as at school-level based on the two-level analyses that were described in

the previous chapter followed closely the variation at the corresponding levels based on these three-level analyses. Thus, as far as the amount of variance available to be explained at the student and at school levels are concerned, it does not seem to matter markedly whether the two-level or three-level analyses are to employed.

Three-level unconditional models

The next step of the analyses is to model achievement in numeracy and literacy as the outcome variables predicted by student-level variables. No predictors are specified at Levels 2 and 3. The step-up approach is followed to examine which of the student-level variables have a significant influence on achievement in numeracy and literacy.

The final Level-1 unconditional models for numeracy and literacy are presented in the Equations provided below for Models X, Y and Z. It should be noted that for Model-X, the final unconditional models for numeracy and literacy are similar and are therefore reported together below.

Model-X

For both numeracy and literacy

$$Y = P0 + P1*(YEARLEVL) + P2*(AGE) + P3*(ATSI) + P4*(HOME) + E \quad \text{Equation 8.7}$$

Model-Y

For numeracy

$$Y = P0 + P1*(SEX) + P2*(TRANS) + P3*(AGE) + P4*(ATSI) + P5*(INOZ) + P6*(Y3NSCORE) + E \quad \text{Equation 8.8}$$

For literacy

$$Y = P0 + P1*(SEX) + P2*(TRANS) + P3*(AGE) + P4*(ATSI) + P5*(HOME) + P6*(INOZ) + P7*(Y3LSCORE) + E \quad \text{Equation 8.9}$$

Model-Z

For numeracy

$$Y = P0 + P1*(SEX) + P2*(AGE) + P3*(ATSI) + P4*(INOZ) + P5*(Y3NSCORE) + E \quad \text{Equation 8.10}$$

For literacy

$$Y = P0 + P1*(SEX) + P2*(AGE) + P3*(ATSI) + P4*(HOME) + P5*(INOZ) + P6*(Y3LSCORE) + E \quad \text{Equation 8.11}$$

At Level-1, the three-level unconditional models for numeracy and literacy presented in the above equations are exactly the same as the two-level unconditional models presented in the previous chapter. This is no surprise because the same data are used for the two-level analyses as are used for the three-level analyses. Furthermore, at Level-1, the structure of the three level models is exactly the same as the structure of the two level models encountered in the previous chapter.

The three-level analyses indicate that at the student-level, the variables that have a significant influence on the outcome variables are:

(a) Grade Level (YEARLEVL), Age of the Student (AGE), Racial Background (ATSI) and Speaking English at Home (HOME) in Model-X;

(b) Sex of the Student (SEX), Age of the Student, Racial Background, Living in Australia (INOZ), Transience (TRANS), Prior Achievement (Y3NSCORE or

Y3LSCORE) and Speaking English at Home (for literacy only) in Model-Y; and

(c) Sex of the Student, Age of the Student, Racial Background, Living in Australia and Prior Achievement (Y3NSCORE or Y3LSCORE) and Speaking English at Home (for literacy only) in Model-Z.

Thus, at the student-level, the results of the three-level analyses are identical to results of the two-level analyses presented in the previous chapter.

The reliability estimates of the student-level variables with random effects at a particular level of the hierarchy are presented in Appendix 14.2. The values of these reliability estimates exceed the minimum value (0.05) recommended by Bryk and Raudenbush (1992).

Final three-level models

As mentioned above, further HLM runs are undertaken to build up the equations at the school-level through adding the significant school-level variables to the equation using the step-up strategy and the exploratory analysis sub-routine. Then the final stage of the three-level analysis is undertaken to build up the model at the occasion-level following the step-by-step procedure and using the exploratory analysis routine as described above. For some models (for example, Model-Y for both numeracy and literacy) there are no potential Level-3 predictors with absolute 't-to-enter' values greater than 1.96, and therefore, no predictors can be entered into these models at Level-3.

The final three-level hierarchical models for numeracy and literacy for the three types of models are presented in Figures 8.1 to 8.6. In the diagrams shown in Figures 8.1 to 8.6, only the factors that have a significant ($p<0.05$) direct (or interaction) effect on student achievement have been displayed. An ‘effect’ is considered to be significant at the $p=0.05$ level if its coefficient taken in absolute terms is more than twice its standard error (given in parenthesis in these diagrams). The coefficients displayed in these diagrams are the ones that are obtained using unstandardized variables.

The final estimations of the fixed effects from the three-level HLM analysis for the models shown in Figures 8.1 to 8.6 are displayed in Tables 8.3 and 8.4 for numeracy and literacy respectively. Both the standardized as well as the metric regression coefficients of the variables included in the final models are presented in the tables.

From Tables 8.3 and 8.4, it should be noted that some variables have their random effect fixed across schools and/or fixed across occasions. This is because the program had problems converging after the inclusion of these variables or the reliability estimates of the variables fell below the 0.05 recommended by Raudenbush and Bryk (1992). For example, for Model-X the program had problems converging after inclusion of the variable AGE and again after inclusion of variable HOME, and therefore, the effects associated with AGE and HOME are fixed across schools. Hox (1994) has argued that if the estimation procedure does not converge, it is an indication that something is wrong. Consequently, Hox has recommended the fixing of the regression slopes of the variables with low reliability to solve the problem.

By and large, the final three-level models at Levels 1 and 2 for numeracy and literacy presented in Tables 8.3 and 8.4 are basically the same as the final two-level models presented in the previous chapter. The similarity is because the same data are used for both analyses with the only difference being that the three-level analysis allows the identity of the school to be kept intact unlike the two-level analysis where different identities are used for the same school on different occasions.

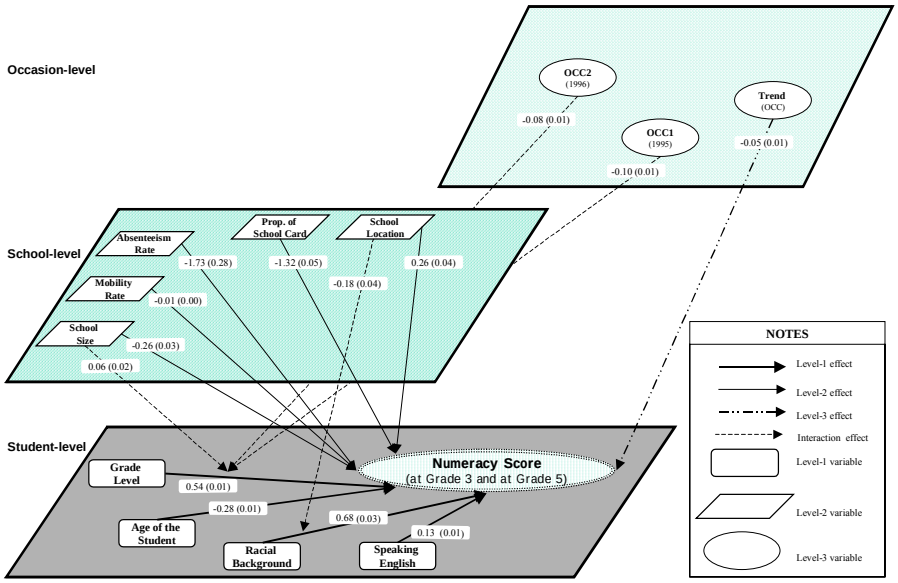


Figure 8.1 Final three-level hierarchical model for numeracy – Model-X

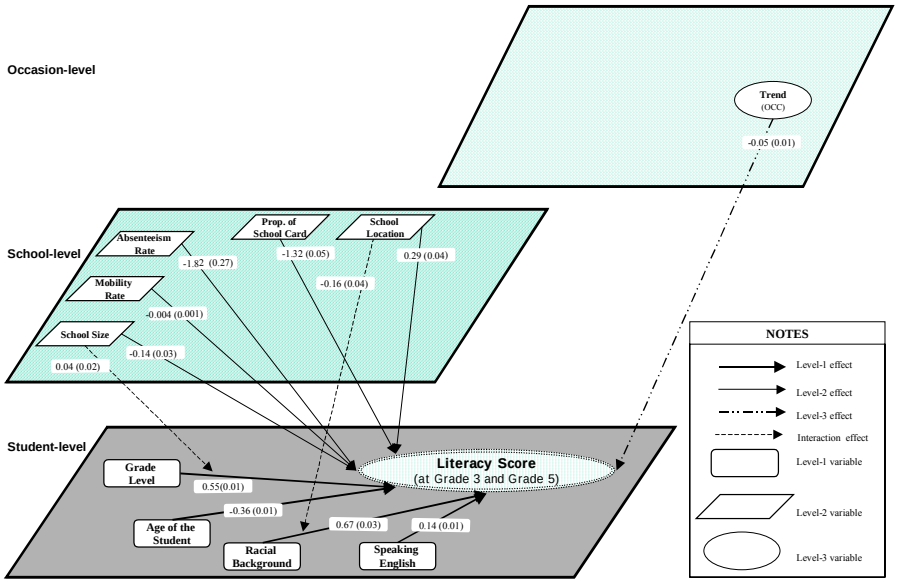


Figure 8.2 Final three-level hierarchical model for literacy – Model-X

In the next three sub-sections, interpretations of the results of the final fixed effects displayed in Tables 8.3 and 8.4 are presented. An interpretation of the results at the student level is presented in the first sub-section, followed by an interpretation of the results at the school level and occasion level in the second and third sub-sections respectively. Because the cross-level interaction effects in the three-level analyses are

basically similar to the ones in the two-level analyses, the interested reader is referred to Chapter 7, where a treatment of these cross-level interaction effects is provided.

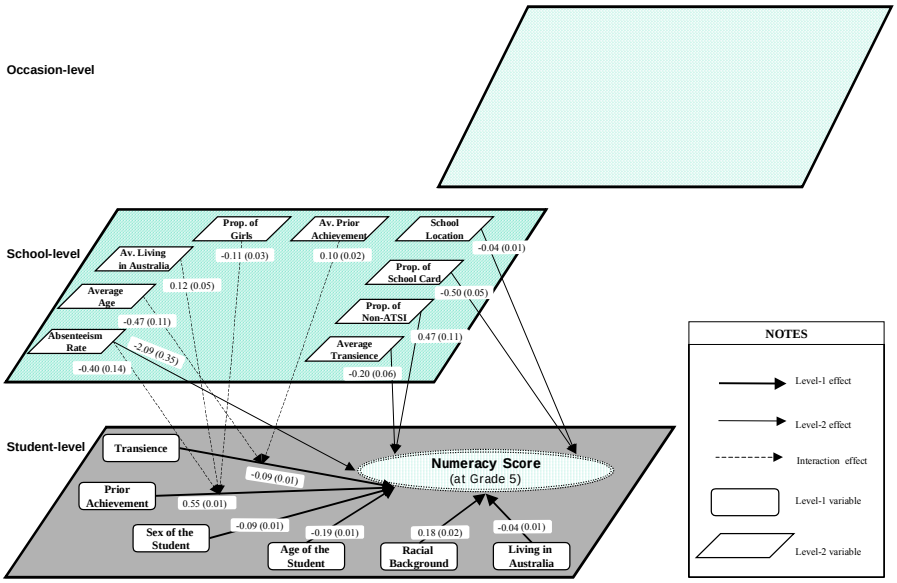


Figure 8.3 Final three-level hierarchical model for numeracy – Model-Y

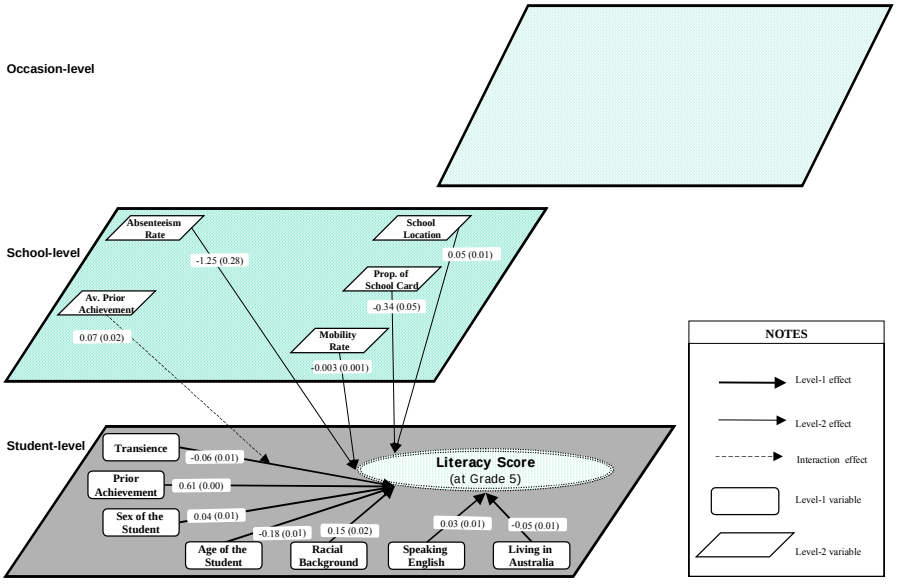


Figure 8.4 Final three-level hierarchical model for literacy – Model-Y

Student-level model

The results of the three-level (in Tables 8.2 and 8.3) and the two-level analyses presented in the previous chapter indicate the same relationships regarding the student-level factors influencing achievement in numeracy and literacy among Grades 3 and 5 students in primary schools in South Australia when other things are equal.

1. Students at Grade 5 are likely to achieve better in numeracy and literacy than students at Grade 3 by about one logit.
2. Younger students are likely to achieve better in numeracy and literacy than their older counterparts.
3. Aboriginal and Torres Strait Islander students are likely to achieve at lower levels in numeracy and literacy than students of other races.

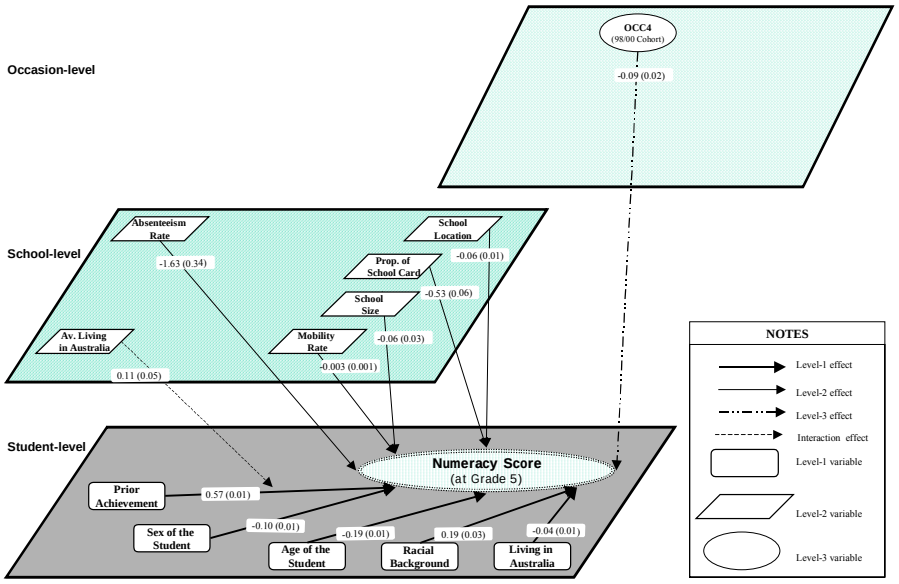


Figure 8.5 Final three-level hierarchical model for numeracy – Model-Z

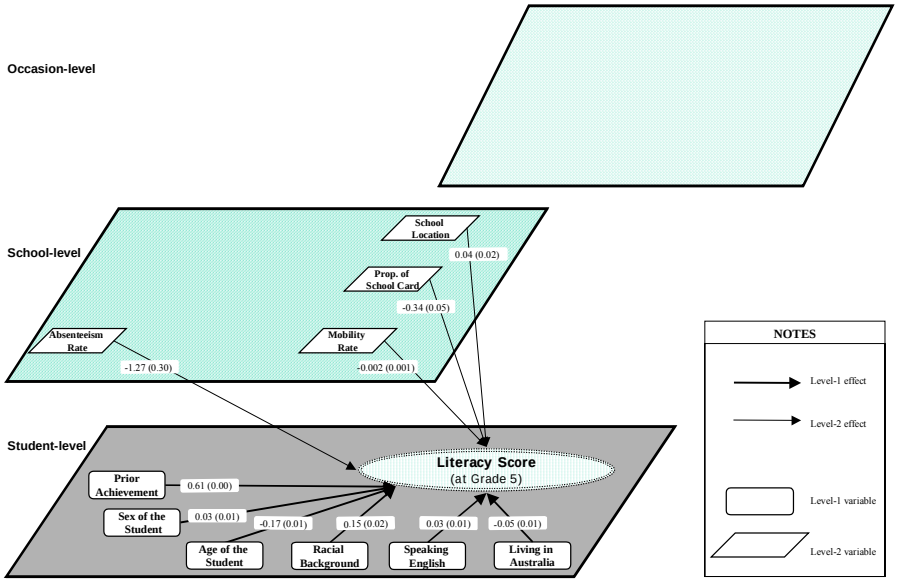


Figure 8.6 Final three-level hierarchical model for literacy – Model-Z

Table 8.2 Final estimation of fixed effects from the final three-level numeracy models

						Model - X					Model - Y					Model - Z				
						Coefficient[ξ]		SE	T-ratio	P-value	Coefficient[ξ]		SE	T-ratio	P-value	Coefficient[ξ]		SE	T-ratio	P-value
						Std'zed	Metric				Std'zed	Metric				Std'zed	Metric			
For		**INTRCPT1,**			**P0**															
	For	INTRCPT2, INTRCPT3,	G000	B00		0.70	-0.51	0.03	-15.65	0.00 *jk*	1.34	1.23	0.03	42.88	0.00 *jk*	1.40	1.28	0.03	48.82	0.00 *jk*
		OCC4,	G004													-0.04	-0.09	0.02	-4.87	0.04
		OCC,	G007			-0.05	-0.05	0.01	-6.02	0.00										
	For	METRO, INTRCPT3,	G010	B01		0.04	0.26	0.04	6.48	0.00										
	For	PSCARD, INTRCPT3,	G020	B02		-0.25	-1.32	0.05	-27.48	0.00	-0.10	-0.50	0.05	-10.45	0.00	-0.10	-0.53	0.06	-9.65	0.00
	For	MOBILITY, INTRCPT3,	G030	B03		-0.10	-0.01	0.00	-5.92	0.00						-0.05	-0.00	0.00	-2.92	0.00
	For	ABSENT, INTRCPT3,	G040	B04		-0.08	-1.73	0.28	-6.23	0.00	-0.10	-2.09	0.35	-6.01	0.00	-0.08	-1.63	0.34	-4.75	0.00
	For	SSIZELOG, INTRCPT3,	G050	B05		-0.08	-0.26	0.03	-8.19	0.00						-0.02	-0.06	0.03	-2.05	0.04
	For	GPOLOG, INTRCPT3,	G060	B06							-0.03	-0.04	0.01	-3.44	0.00	-0.04	-0.06	0.01	-4.29	0.00
	For	ATSI_1, INTRCPT3,	G070	B07							0.07	0.47	0.11	4.42	0.00					
	For	TRANS_1, INTRCPT3,	G080	B08							-0.04	-0.20	0.06	-3.55	0.00					
For		**SEX,**			**P1**															
	For	INTRCPT2, INTRCPT3,	G100	B10		xxx	xxx	xxx	xxx	xxx	-0.05	-0.09	0.01	-11.68	0.00	-0.05	-0.10	0.01	-11.91	0.00
For		**AGE,**			**P2**															
	For	INTRCPT2, INTRCPT3,	G200	B20		-0.11	-0.28	0.01	-37.04	0.00	-0.07	-0.19	0.01	-16.87	0.00	-0.07	-0.19	0.01	-15.24	0.00
For		ATSI,			**P3**															
	For	INTRCPT2, INTRCPT3,	G300	B30		0.11	0.68	0.03	23.94	0.00 *j*	0.03	0.18	0.02	7.68	0.00	0.03	0.19	0.03	7.68	0.00
	For	METRO, INTRCPT3,	G310	B31		-0.02	-0.18	0.04	-4.94	0.00										
For		**HOME,**			**P4**															
	For	INTRCPT2, INTRCPT3,	G400	B40		0.08	0.13	0.01	26.20	0.00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
For		**INOZ,**			**P5**															
	For	INTRCPT2, INTRCPT3,	G500	B50		xxx	xxx	xxx	xxx	xxx	-0.01	-0.04	0.01	-3.29	0.00	-0.01	-0.04	0.01	-2.99	0.00
For		**TRANS,**			**P6**															
	For	INTRCPT2, INTRCPT3,	G600	B60							-0.03	-0.09	0.01	-7.01	0.00 *j*					
	For	Y3NSCO_1, INTRCPT3,	G610								0.03	0.10	0.02	4.60	0.00					
	For	AGE_1, INTRCPT3,	G620								-0.03	-0.47	0.11	-4.14	0.00					
For		**YEARLEVL,**			**P7**															
	For	INTRCPT2, INTRCPT3,	G700	B70		0.51	0.54	0.01	82.44	0.00 *jk*										
		OCC1,	G701			-0.03	-0.10	0.01	-7.52	0.00										
		OCC2,	G702			-0.03	-0.08	0.01	-5.59	0.00										
	For	SSIZELOG, INTRCPT3,	G710	B71		0.02	0.06	0.02	3.53	0.00										
For		**Y3NSCORE,**			**P8**															
	For	INTRCPT2, INTRCPT3,	G800	B80							0.69	0.55	0.01	121.24	0.00	0.70	0.57	0.01	116.52	0.00 *j*
	For	ABSENT, INTRCPT3,	G810	B81							-0.02	-0.40	0.14	-2.94	0.00					
	For	SEX_1, INTRCPT3,	G820	B82							-0.02	-0.11	0.03	-3.08	0.00					
	For	INOZ_1, INTRCPT3,	G830	B83							0.02	0.12	0.05	2.59	0.01	0.01	0.11	0.05	2.07	0.04

Notes: xxx - Variable has no significant influence on the outcome.
Shade - Variable not available for examination in this model.
k - Residual parameter of this coefficient is left to vary at the occasion-level.
ξ - Standard errors (SE), t-ratios and p-values presented are those obtained using unstandardized variables.
j - Residual parameter of this coefficient is left to vary at the school-level.

Table 8.3 Final estimation of fixed effects from the final three-level literacy models

						Model - X					Model - Y					Model - Z				
						Coefficient[ξ]		SE	T-ratio	P-value	Coefficient[ξ]		SE	T-ratio	P-value	Coefficient[ξ]		SE	T-ratio	P-value
						Std'zed	Metric				Std'zed	Metric				Std'zed	Metric			
For		**INTRCPT1,**			**P0**															
	For	INTRCPT2, INTRCPT3,	G000	B00		0.77	-0.49	0.04	-13.41	0.00 *jk*	1.47	1.29	0.06	20.23	0.00 *jk*	1.51	1.33	0.07	20.34	0.00 *jk*
		OCC,	G007			-0.08	-0.05	0.01	-3.90	0.03										
	For	METRO, INTRCPT3,	G010	B01		0.07	0.29	0.04	7.13	0.00	0.02	0.05	0.01	3.16	0.00	0.02	0.04	0.02	2.89	0.00
	For	PSCARD, INTRCPT3,	G020	B02		-0.25	-1.32	0.05	-29.18	0.00	-0.07	-0.34	0.05	-7.21	0.00	-0.06	-0.34	0.05	-6.74	0.00
	For	MOBILITY, INTRCPT3,	G030	B03		-0.08	-0.00	0.00	-4.82	0.00	-0.05	-0.00	0.00	-3.59	0.00	-0.04	-0.00	0.00	-2.81	0.01
	For	ABSENT, INTRCPT3,	G040	B04		-0.09	-1.82	0.27	-6.83	0.00	-0.06	-1.25	0.28	-4.39	0.00	-0.06	-1.27	0.30	-4.27	0.00
	For	SSIZELOG, INTRCPT3,	G050	B05		-0.04	-0.14	0.03	-4.84	0.00										
For		**SEX,**			**P1**															
	For	INTRCPT2, INTRCPT3,	G100	B10		×××	×××	×××	×××	×××	0.02	0.04	0.01	5.20	0.00	0.02	0.03	0.01	3.91	0.00
For		**AGE,**			**P2**															
	For	INTRCPT2, INTRCPT3,	G200	B20		-0.14	-0.36	0.01	-46.93	0.00	-0.06	-0.18	0.01	-17.55	0.00	-0.06	-0.17	0.01	-15.86	0.00
For		**ATSI,**			**P3**															
	For	INTRCPT2, INTRCPT3,	G300	B30		0.11	0.67	0.03	22.46	0.00 *j*	0.03	0.15	0.02	7.59	0.00	0.03	0.15	0.02	6.56	0.00
	For	METRO, INTRCPT3,	G310	B31		-0.02	-0.16	0.04	-4.14	0.00										
For		**HOME,**			**P4**															
	For	INTRCPT2, INTRCPT3,	G400	B40		0.09	0.14	0.01	28.79	0.00	0.02	0.03	0.01	3.89	0.00	0.01	0.03	0.01	3.12	0.00
For		**INOZ,**			**P5**															
	For	INTRCPT2, INTRCPT3,	G500	B50		×××	×××	×××	×××	×××	-0.02	-0.05	0.01	-5.12	0.00	-0.02	-0.05	0.01	-4.89	0.00
For		**TRANS,**			**P6**															
	For	INTRCPT2, INTRCPT3,	G600	B60							-0.02	-0.06	0.01	-5.08	0.00 *j*					
	For	Y3LSCO_1, INTRCPT3,	G610								0.02	0.07	0.02	3.18	0.00					
For		**YEARLEVL,**			**P7**															
	For	INTRCPT2, INTRCPT3,	G700	B70		0.55	0.55	0.01	104.88	0.00 *j*										
	For	SSIZELOG, INTRCPT3,	G710	B71		0.01	0.04	0.02	2.55	0.01										
For		**Y3LSCORE,**			**P8**															
	For	INTRCPT2, INTRCPT3,	G800	B80							0.77	0.61	0.00	197.66	0.00 *j*	0.77	0.61	0.00	187.90	0.00

Notes: ××× - Variable has no significant influence on the outcome.
Shade - Variable not available for examination in this model.
k - Residual parameter of this coefficient is left to vary at the occasion-level.
ξ - Standard errors (SE), t-ratios and p-values presented are those obtained using unstandardized variables.
j - Residual parameter of this coefficient is left to vary at the school-level.

4. Students who always speak English at home are likely to achieve better in numeracy (Model-X only) and literacy (all the Models) than students who rarely (or never) speak English at home.
5. Boys are likely to achieve better in numeracy than girls, while girls are likely to achieve better than boys in literacy.
6. Students who are new to Australia are likely to achieve better in numeracy and literacy than the students who were born in Australia.
7. Students who remain in the same school over the two-year duration are likely to achieve better in numeracy and literacy than students who change schools.
8. Students who have high scores on the BST at Grade 3 are likely to perform better on the tests at Grade 5 than students who have low scores on the tests at Grade 3.

School-level model

At the school-level, the results of the three-level analyses also agree with the results of the two-level analyses. Thus, when all other things are equal, the following can be said regarding school-level factors influencing student achievement in numeracy and literacy among Grades 3 and 5 students in primary schools in South Australia.

1. Students in schools with lower proportions of school-card holders are likely to achieve better in numeracy and literacy than students in schools with high proportions of school-card holders.
2. Students in schools with low mobility (or transience) rates are likely to achieve better in numeracy and literacy than students in schools with high mobility (or transience) rates.
3. Students in schools with low rates of absenteeism are likely to achieve better in numeracy and literacy than students in schools with high rates of absenteeism.
4. Students in schools that are located in urban areas (or near Adelaide) are likely to perform better than students in schools that are located in rural areas (or far from Adelaide).
5. Students in schools with fewer students are likely to achieve better in numeracy and literacy than students in schools with many students (Model-X). However, the results of the analyses from Model-Y indicate that the size of the school does not influence achievement on the BST while the results from Model-Z indicate that the size of the school influences achievement in numeracy but not in literacy.
6. Students in schools with high proportions of non-ATSI students are likely to achieve better in numeracy (but not necessarily in literacy) than students in schools with high proportions of Aboriginal and Torres Strait Islander students (Model-Y only).

Again, of the 12 variables examined at the school-level, seven had no direct significant influence on achievement in numeracy and literacy in any of the models examined. The seven variables are namely Proportion of Girls (SEX_1), Average Age (AGE_1), Average Speaking English (HOME_1), Average Living in Australia (INOZ_1), Country Area Program (CAP), Participation Size (YR35PPT) and Prior Achievement (Y3NSCO_1 for numeracy or Y3LSCO_1 for literacy). Three out of these seven variables mentioned here; namely CAP, YR35PPT and HOME_1 have no interaction effects with any of the variables in any of the models examined.

Occasion-level model

At Level-3 of the models, the results in Tables 8.2 and 8.3 seem to disagree with the results of the two-level analyses presented in the previous chapter in several ways.

First, the results of the three-level analyses in Tables 8.2 and 8.3 indicate that the linear trend variable (OCC) has a significant influence on achievement in numeracy and literacy at the occasion-level in Model-X but not in the other two types of models. The negative metric coefficients and negative t-ratio values for OCC on the intercept for numeracy (-0.05, t=-6.02) and for literacy (-0.05, t=-3.90) in Model-X indicates that there are significant downward trends in the grand mean between occasions. On the other hand, the results for Models Y and Z indicate the linear trend variable OCC does not show a significant influence on achievement in numeracy and literacy. These results seem to be contrary to those of the two-level analyses that indicate that the variable OCC has a significant influence on achievement in numeracy and literacy in Models Y and Z but not in Model-X. However, in this case, the results of the three-level analyses are more appropriate than those obtained from the two-level analyses because in the three-level analyses schools are linked over time. Nevertheless, for the three-level analyses, it should be remembered that the number of units at the third level (that is, the occasion-level) is too small (only six units for Model-X and only four units for Models Y and Z) for sound significance testing at that level.

Second, disagreeing with the results of the two-level analyses, the results in Table 8.2 indicate that the dummy variable OCC4 has a significant influence on achievement in numeracy in Model-Z at the occasion-level. The negative metric coefficient and t-ratio values indicate that the average performance of the schools on the fourth occasion (that is, performance of the schools based on the 1998/2000 Cohort) was lower than on the other three occasions (-0.09, t=-4.87). The results in Tables 8.2 and 8.3 also show that none of the other dummy variables representing the testing occasions (or student cohorts) have direct and significant influences on achievement in numeracy and literacy in any of models examined. These results are different from the results of the two-level analyses where (a) the variable OCC6 (that is, 2000) in Model-X has a significant influence in literacy, and (b) the variable OCC3 (the 1997/1999 cohort) in Models Y and Z has significant influences on achievement in numeracy and literacy. Here again it should be taken into consideration that the results of the three-level analyses are likely to be more appropriate compared with those of the two-level analyses, but the difficulty associated with significance testing in the three-level analyses at the third level of the models should be borne in mind.

Finally, the results in Tables 8.2 and 8.3 also differ from the results of the two-level analyses regarding the dummy variables denoting the six occasions that have significant influences on the Grade Level (YEARLEVL) slope for numeracy and literacy. For numeracy, results of the three-level analyses indicate that in 1995 (OCC1; -0.10, t=-7.52) and again in 1996 (OCC2; -0.08, t=-5.59) the growth in achievement was lower than on the other occasions while the results for the two-level analyses indicate that the growth in achievement is lower in 1995 (OCC1; -0.04, t=-4.09) only. For literacy, the results of the three-level analyses indicate that growth in achievement remains almost the same between occasion because none of the dummy variables have a significant influence on the YEARLEVL slope. On the other hand, the results of the two-level analyses indicate that the growths in achievement in literacy are significantly lower in 1995 (OCC1, -0.07, t=-6.36) and 1999 (OCC5, -0.13, t=-14.02) but significantly higher in 1997 (OCC3, 0.03, t=2.74).

It should be emphasized that the above differences between the significance of the occasion variables in the three-level models and in the two-level models should be

interpreted with caution. Of crucial importance is the recognition that the differences are a consequence of the difference between the multilevel structures of the three-level and two-level models, which differ in the number of units involved in significance testing. Therefore, the differences noted above do not necessarily indicate contradictions in the findings.

Estimation of variance explained

The results of the final estimations of the variance components for the final three-level models and the results of the variance components obtained from the null models (provided in Table 8.1) are presented together in Table 8.4 (in rows 'a' and 'a') for ease of comparison. From the information presented in rows 'a' and 'b', the information presented in rows 'c' to 'f' are calculated. A discussion of the calculations involved here is presented in Chapter 4.

For example, for literacy, the predictors included in the final transience model (Model-Y), explain 54.4 per cent of the 81.9 per cent variance available at the student-level, and that is equal to 44.6 per cent (that is, 54.4×81.6) of the total variance explained at the student-level. Similarly, for the same model, the predictors included in the final model explain 11.7 per cent (that is, 71.8 per cent of 16.3 per cent) at the school-level, and explains 0.6 per cent (that is, 35.4 per cent of 1.8 per cent) at the occasion-level. Therefore, the total variance explained by the predictors included in the final three-level transience model for literacy is $44.6 + 11.7 + 0.6 = 56.9$, which leaves 43.1 per cent of the total variance unexplained.

Thus, the results in Table 8.4 show that the percentages of variance that are left unexplained (in the shaded cell of row 'f') in Model-X are much larger compared to the percentages of variance that are left unexplained in either Model-Y or Model-Z. In addition, the percentages of variances left unexplained in Model-Y follow closely the percentages of variance that are left unexplained in Model-Z for numeracy as well as literacy.

However, for all three models examined here, the results in Table 8.4 show that the percentages of variance that are left unexplained at the student-level are higher when compared with the percentages of variances that are left unexplained at the school-level (given in **bold** in row 'f') and the occasion-level. The results also indicate that almost negligible percentages of variance are left unexplained at the occasion-level in the final models.

For Models Y and Z, it can also be noted from the results in Table 8.4 that the amounts of variance explained in the final models are noticeably large especially at the school-level where approximately 70 per cent or more of the available variance is explained. Consequently, only around five per cent of the total amounts of variance at the school-level is left unexplained. From these results, it can further be noted that the total variances that are left unexplained (in the shaded cell of row 'f') for numeracy are noticeably larger compared to the total amounts of variance that are left unexplained for literacy.

Comparison of model fit using the deviance statistic

HLM5/3L computes the deviance statistic for the model tested together with the number of parameters in the model for each run just as was described for HLM5/2L in the previous chapter.

Table 8.4 Estimation of variance explained using the three-level numeracy and literacy models

	Model - X [a]				**Model - Y** [b]				**Model - Z** [c]			
	Level-1	Level-2	Level-3	Total	Level-1	Level-2	Level-3	Total	Level-1	Level-2	Level-3	Total
	(N=14,4346)	(N=2,868)	(N=6)		(N=37,832)	(N=1,853)	(N=4)		(N=32,741)	(N=1,823)	(N=4)	
Numeracy												
a) Var. Comp. Null Model	1.18	0.39	0.00	1.57	1.00	0.22	0.00	1.22	0.99	0.21	0.00	1.20
b) Var. Comp. Final Model	1.14	0.33	0.00		0.56	0.06	0.00		0.54	0.06	0.00	
c) Var. Available	75.0%	24.8%	0.2%		81.8%	18.1%	0.2%		82.2%	17.6%	0.2%	
d) Var. Explained	2.7%	15.6%	75.7%		44.5%	71.8%	48.1%		45.3%	70.4%	94.8%	
e) Total Var. Explained	2.0%	3.9%	0.1%	6.0%	36.4%	13.0%	0.1%	49.5%	37.3%	12.4%	0.2%	49.8%
f) Var. Left Unexplained	73.0%	**21.0%**	0.0%	94.0%	45.3%	**5.1%**	0.1%	50.5%	44.9%	**5.2%**	0.0%	50.2%
Literacy												
a) Var. Comp. Null Model	1.18	0.39	0.00	1.57	1.00	0.20	0.02	1.22	0.98	0.19	0.02	1.19
b) Var. Comp. Final Model	1.17	0.29	0.00		0.46	0.06	0.01		0.44	0.06	0.01	
c) Var. Available	75.0%	24.8%	0.2%		81.9%	16.3%	1.8%		81.8%	16.3%	1.9%	
d) Var. Explained	0.7%	26.4%	21.0%		54.4%	71.8%	35.4%		54.6%	71.4%	34.2%	
e) Total Var. Explained	0.5%	6.6%	0.0%	7.1%	44.6%	11.7%	0.6%	56.9%	44.7%	11.6%	0.6%	57.0%
f) Var. Left Unexplained	74.5%	**18.3%**	0.1%	92.9%	37.4%	**4.6%**	1.2%	43.1%	37.1%	**4.7%**	1.2%	43.0%

Note: a - For Model-X, the simplest model has the variable YEARLEVL as the only predictor.
b - Transience model
c - Non-transience model

However, using HLM5/3L the MLR method is not available for solution estimation of three-level models and, therefore, the MLF procedure has to be used.

For each of the HLM runs, the chi-square test is used to compare the fit of a model with the preceding model. The steps undertaken in the three-level model comparison are similar to the ones undertaken in the two-level model comparison. That is, the optional hypothesis testing sub-routine is employed to compare model fit in successive HLM runs by entering the deviance statistic and number of parameters reported in the output file of a previous model into the optional hypothesis testing dialog box fields provided in HLM5/3L. A chi-square statistic, with associated degrees of freedom and p-value are then printed at the end of the next HLM5/3L output file.

Table 8.5 presents results of deviance statistics and the chi-square tests carried out to compare model fit at the conclusion of the unconditional part and at the conclusion of the final models for numeracy and literacy for the three types of models examined. In Table 8.5, the fit of the unconditional model is compared to the fit of the null model (or grade-level-only for Model-X), and the fit of the final model is compared to the fit of the unconditional model.

The chi-squares tests presented in Table 8.3 indicate better fit of the unconditional model compared to the null (or grade-level-only) model and better fit of the final model compared to the unconditional model for all the types of models examined. Therefore, the inclusion of the predictors at the three levels of hierarchy significantly improves the overall fit of the models.

Conclusions

The results of the three-level HLM analyses closely match the results of the two-level HLM analyses regarding the student-level and school-level variables that have significant influences on achievement in numeracy and literacy. However, the results of the three-level analyses vary in some ways with the results of the two-level analyses regarding the general linear trend in the mean achievement of the schools in the State over time.

However, the differences between the significance of the linear trend variable in the three-level models and in the two-level models do not necessarily indicate contradictory findings. This is because the differences are arguably a consequence of the difference between the multilevel structures of the three-level and two-level models. Nevertheless, the results of the three-level analyses should in this case be considered superior to the results of the two-level analyses because the three-level structure recognizes that schools are coherent entities which persist (Paterson, 1991). That is, even though separate cross-sections of Grades 3 and 5 students are taken on each occasion, the three-level structure allows the identity of the school to be kept intact, unlike the structure employed in the two-level analyses where different identities are used for the same school on different occasions. Paterson (1991) has argued that keeping the identity of the school intact disentangles true change from sampling error, and therefore, provides a clearer idea of how the system is changing. Therefore, the three-level analyses reported in this chapter should provide a better picture of the performance of the schools over time than the two-level analyses reported in the previous chapter.

Despite having said that the three-level analyses should provide a better picture of performance of the schools over time than the two-level analyses, it should be remembered that the number of units at the third level is too small for sound significance testing at that level. Obviously, based on either the two-level or three-

level analyses reported this far, it is not clear what could be happening to the performance of the schools over time. A better picture of the performance of the schools over time is provided in the next chapter by employing three-level longitudinal models that allow the identify of the schools to be kept intact while at the same time keeping the number of units at the school-level large.

Table 8.5 Comparison of model fit using the chi-square tests

a) Model-X

	Deviance Statistic	Number of Parameters	Chi-square Statistic	Degrees of Freedom	P-value
Numeracy					
Grade-level-only	441,827.42	9			
Unconditional	437,399.15	15	4,428.27	6	0.00
Final	435,955.59	25	1,443.56	10	0.00
Literacy					
Grade-level-only	445,016.18	9			
Unconditional	439,555.33	15	5,460.85	6	0.00
Final	438,217.90	21	1,337.44	6	0.00

b) Model-Y

	Deviance Statistic	Number of Parameters	Chi-square Statistic	Degrees of Freedom	P-value
Numeracy					
Null	110,148.94	4			
Unconditional	88,321.97	15	21,826.97	11	0.00
Final	87,814.92	25	507.05	10	0.00
Literacy					
Null	109,975.40	4			
Unconditional	80,039.69	16	29,935.72	12	0.00
Final	79,997.97	18	41.71	2	0.00

c) Model-Z

	Deviance Statistic	Number of Parameters	Chi-square Statistic	Degrees of Freedom	P-value
Numeracy					
Null	473,041.39	4			
Unconditional	437,399.15	15	35,642.24	11	0.00
Final	435,963.87	22	1,435.28	7	0.00
Literacy					
Null	478,616.75	4			
Unconditional	439,555.33	15	39,061.42	11	0.00
Final	438,217.90	21	1,337.44	6	0.00

For all three types of models, the results of the three-level analyses show that in terms of achievement in numeracy and literacy, there are (a) huge variations between students within schools compared to the variation in performance between schools; and (b) very little (less than 2.0 per cent) variance in schools between occasions. The results also indicate that, in the final models, the total variances explained at Level-1

are generally low compared to the variance available at that level. However, the amounts of variance explained at Levels 2 and 3 are high especially in Models Y and Z where very small percentages of variance (around 5%, and mostly less than 1% for Levels 2 and 3 respectively) are left unexplained at the two levels of the hierarchy.

Table 8.6 gives comparisons of the amounts of variances available, explained, and left unexplained at the student-level, school-level and in all the levels combined that were estimated using the two-level analyses and those that were estimated using the three-level analyses for numeracy and literacy. The information provided in Table 8.6 shows that the results of the three-level analyses overwhelmingly agreed with the results of the two-level analyses regarding the amounts of variance available, explained and consequently the amounts left unexplained at the student-level and school-level, as well as in all the levels combined. For all three types of models proposed, the information in Table 8.6 shows that the amounts of variance left unexplained when either the two-level analyses or the three-level analyses are applied remains almost the same. Obviously, the three-level analyses offer no added advantage as far as the amounts of variance explained in the final models are concerned.

The information in Table 8.6 also show that regardless of the type of analyses that are employed, the variances left unexplained at the school-level are small in Models Y and Z where Prior Achievement at the student-level is taken into account. This is despite the fact that several student-level variables, such as socioeconomic status and grade repetition, which have been shown in other studies to influence academic achievement, are not available for examination in this study.

Potential implications

Without doubt, the discussions and analyses presented in the previous chapter and in this chapter show that, after taking into account achievement in the BST at Grade 3, a very small amount of the variance available in the Grade 5 scores is left unexplained at the school-level. In addition, the analyses show that controlling for Prior Achievement alone is enough to reduce the variance at the school-level substantially. Moreover, the analyses show that the variance left unexplained at the school-level by controlling for Prior Achievement alone is almost equal to that left unexplained when all student-level and school-level (and where applicable occasion-level) factors are included in the analyses. These results are consistent, whether all students who could be matched are considered (transience model), or whether those who remain in the same school over the two-year period are considered (non-transience model), and consistent, whether two-level analyses are employed, or whether three-level analyses are employed.

The results are also consistent across the two subject areas included in the BST, namely, numeracy and literacy. These findings have potential implications for research into school effects, especially if scores from the BST are to be used as inputs for computation of the indicators of school performance across the two grade levels.

First, it is logical to question the appropriateness of computing school performance indicators using students' scores obtained from the BST on one occasion only, given that such indicators could end up being used to compare or rank schools. Arguably, such comparison or ranking of schools would rely on the variance left unexplained after accounting for either:

(a) student-level factors that have a significant influence on achievement, that is, a Type A effect (Willms and Raudenbush, 1989; Harker and Nash, 1996); or

Table 8.6 Estimates variances using the two-level analyses and using the three-level analyses

	Model-X			Model-Y			Model-Z		
	Student-level	School-level	All levels	Student-level	School-level	All levels	Student-level	School-level	All levels
Numeracy									
a) Two-level analyses									
Var. Available	75.1%	24.9%		81.7%	18.3%		82.1%	17.9%	
Var. Explained	2.9%	17.1%		44.3%	72.1%		45.2%	70.7%	
Total Var. Explained	2.2%	4.3%	6.4%	36.2%	13.2%	49.4%	37.1%	12.6%	49.7%
Var. Left Unexplained	72.9%	**20.6%**	93.6%	45.5%	**5.1%**	50.6%	45.0%	**5.2%**	50.3%
b) Three-level analyses									
Var. Available	75.0%	24.8%		81.8%	18.1%		82.2%	17.6%	
Var. Explained	2.7%	15.6%		44.5%	71.8%		45.3%	70.4%	
Total Var. Explained	2.0%	3.9%	6.0%	36.4%	13.0%	49.5%	37.3%	12.4%	49.8%
Var. Left Unexplained	73.0%	**21.0%**	94.0%	45.3%	**5.1%**	50.5%	44.9%	**5.2%**	50.2%
Literacy									
a) Two-level analyses									
Var. Available	79.0%	21.0%		81.8%	18.2%		81.7%	18.3%	
Var. Explained	3.5%	17.7%		54.6%	75.6%		54.7%	75.1%	
Total Var. Explained	2.7%	3.7%	6.5%	44.6%	13.8%	58.4%	44.7%	13.8%	58.5%
Var. Left Unexplained	76.3%	**17.3%**	93.5%	37.1%	**4.4%**	41.6%	37.0%	**4.6%**	41.5%
b) Three-level analyses									
Var. Available	75.0%	24.8%		81.9%	16.3%		81.8%	16.3%	
Var. Explained	0.7%	26.4%		54.4%	71.8%		54.6%	71.4%	
Total Var. Explained	0.5%	6.6%	7.1%	44.6%	11.7%	56.9%	44.7%	11.6%	57.0%
Var. Left Unexplained	74.5%	**18.3%**	92.9%	37.4%	**4.6%**	43.1%	37.1%	**4.7%**	43.0%

Note: All levels - total for Levels 1 and 2 for the two-level models, and total for Levels 1, 2 and 3 for the three-level models

(b) student-level plus student-related school-level factors that have significant influence on achievement, that is, a Type B effect (Willms and Raudenbush, 1989; Harker and Nash, 1996).

Second, if a school is to be assessed in terms of the value added to student achievement over a one or a two-year period, it would also seem necessary to allow for the performance of the students, before the commencement of the period under review. Consequently, the assessment of the school would rely on the variance left unexplained after accounting for the effects of Prior Achievement on the final achievement at the end of the period, rather than rely on the analyses of scores obtained on one occasion only.

Third, it must be asked whether it is appropriate to rank schools based on the small variance left unexplained. If so, how accurate or how reliable would such comparison or ranking of schools be?

By definition, Type A effect indicators computed using the scores from the BST would reflect the contribution that a given school would make to the increase in achievement in either numeracy or literacy of a particular student if all school-level factors were to remain the same (Harker and Nash, 1996). On the other hand, the Type B effect indicators would reflect the contribution a given school would make to the increase in achievement in either numeracy or literacy of particular students if all student-level factors and school-level factors that are external (see Meyer, 1996; p. 202) to the school were to be taken into account. Subsequently, parents in South Australia choosing a school for their children would be interested in the Type A effect indicator while the general public in South Australia could use the Type B indicators to hold schools accountable for their performance. But how reliable or useful would the information provided by these indicators be to the parents or to those in a position to hold schools accountable given that only a very small amount of variance unexplained is left between schools after controlling for Prior Achievement?

Although there are no simple answers to the above questions, it is obvious that after controlling for Prior Achievement very few differences exists between the primary schools in South Australia that would warrant any comparison or ranking being made using the scores from the BST simply because the amount of unexplained variance is very small at the school-level. Consequently, the school performance indicators computed using the scores from the BST would need to be interpreted with great caution.

Fourth, some critics could argue that in measuring school performance across the two grade levels, it is inappropriate to consider Grade 3 score as assessing Prior Achievement. This is because the student's achievement at the Grade 3 has the school's contribution already embedded in it from Years 1 and 2. However, if the focus of the analysis were to measure the value added by a school to student achievement across the two grade levels, then adjustment for Prior Achievement at Grade 3 would seem appropriate. The question being asked in the analysis would be: how much has the school contributed to the student's achievement since Grade 3?

Schools could be ranked according to the extent of the contribution of value added to the student achievement, but if the residual variance were small, the ranking assigned to schools would be unstable. Such unstable ranking could vary considerably from year to year. Alternatively, a good school might increase its ranking steadily and a poor school might decrease its ranking over time. However, a good school cannot advance beyond a high ranking and a bad school can not drop below a very low ranking. Clearly, based on small residual variance, relative performance is not enough to assess the value added by a school and some measure of absolute performance must

be sought. An approach to measuring school performance that takes into consideration the time that a student takes to learn certain numeracy (or literacy) skills, would seem to be of potential usefulness for primary schools in South Australia. This approach is demonstrated in Chapter 11.

In the next chapter, the approach of estimating school effects and their stability over time proposed by Willms and Raudenbush (1989) is examined. This approach is employed to study the performance of the primary schools in South Australia over time using the scores from the BSTP.

9

Types A and B School Effects

In the two preceding chapters it is shown that the variance left unexplained at the school-level after controlling for factors influencing student performance is small. Consequently, it is argued that value-added ranks assigned to schools based on BSTP data from a single cohort of students are unstable and therefore unreliable and can be misleading.

Thus, the current chapter explores an approach to examining performance of primary schools in South Australia over time using the scores from the BSTP on several cohorts of student. The longitudinal structure that was employed in the models used by Willms and Raudenbush (1989) to estimate school effects and to study their stability is adopted here. A general description of this longitudinal structure is found in Chapter 5. Briefly, this longitudinal structure allows estimation of school effects in terms of a stable index that shows the average effectiveness of the school over the study period, and a change index that shows whether the school had improved or deteriorated in its effectiveness over the study period. Furthermore, the structure allows for the examination of (a) the stability of the school effects, (b) the slope of performance change, and (c) the factors that influence the performance slope.

The first four sections in this chapter describe the specific models used to estimate the school effects based on the longitudinal structure mentioned above, then describe the estimation of the indices of school effectiveness. For both numeracy and literacy, the longitudinal structure is employed to estimate both the stable and change indices of school effectiveness. Two types of school effects, Type A and Type B (Raudenbush and Willms, 1995), are computed for each index of school effectiveness using two data sets; transience and non-transience. The transience data set involves all the students who were matched (N=37,832), while the non-transience data set involves all the matched students who remained in the same school between Grades 3 and 5 (N=32,741). Both types of school effects are estimated using the subtraction method. The theories involved in the estimation of Types A and B school effects are introduced in Chapter 4.

The fifth section of the chapter presents results of the multilevel analyses with attention being paid to the reliability estimates of the components of school effects and the percentages of variance explained at the school-level in the final models for the

estimation of Type A and Type B school effects. The last three sections focus on the correlations among the schools effects across outcome measures and across occasions.

Specification of Type A effects model

It should be remembered that although the general longitudinal structure employed in this chapter has three-levels: its hierarchical nature is different from the general three-level models described in the preceding chapter. The Level-1, Level-2 and Level-3 for the longitudinal structure are student, occasion and school, whereas for the three-level models described in the preceding chapter Level-1, Level-2 and Level-3 were student, school and occasion.

For the current study, and following the notations and arguments presented by Willms and Raudenbush (1989), the three-level model for the estimation of Type A effects based on a longitudinal structure, can be described as follows.

Level-1 model

At the micro-level, the student achievement is modelled as a function of a school mean, student-level background variables plus a random error:

$$\mathbf{Y}_{itj} = \boldsymbol{\pi}_{0tj} + \boldsymbol{\pi}_{htj}\boldsymbol{X}_{hitj} + \mathbf{e}_{itj} \qquad \text{Equation 9.1}$$

where:

$\mathbf{Y}_{itj}$ is the achievement (Rasch score) of student i in school j at occasion t;

$\boldsymbol{\pi}_{0tj}$ is the mean achievement of school j at occasion t;

$\boldsymbol{X}_{hitj}$ are the background characteristics of student i in school j and at occasion t;

$\boldsymbol{\pi}_{htj}$ are the regression coefficients associated with the student background characteristics of school j at occasion t; and

$\mathbf{e}_{itj}$ is a random error or student effect, that is, the deviation of the student mean from the school mean score.

The indices i, j, and t denote students, schools and occasions. There are

$i = 1, 2, \ldots, n_{tj}$ students within school j at occasion t;

$j = 1, 2, \ldots, J_t$ schools for occasion t; and

$t = 1, 2, \ldots, T$ occasions (or student cohorts).

For parsimony, $\boldsymbol{\pi}_{htj}\boldsymbol{X}_{hitj}$ in Equation 9.1 represents the control for several relevant independent variables ($\boldsymbol{\pi}_{1tj}\boldsymbol{X}_{1itj} + \boldsymbol{\pi}_{2tj}\boldsymbol{X}_{2itj} + \ldots + \boldsymbol{\pi}_{htj}\boldsymbol{X}_{hitj}$) that describe student's background characteristics. There are $h = 1, 2, \ldots, H$ independent variables which describe student's background characteristics.

Hence, for the current study, $\boldsymbol{X}_{hitj}$ represents a combination of any of the following student-level variables: Prior Achievement (Y3NSCORE or Y3LSCORE), Sex of the Student (SEX), Age of the Student (AGE), Racial Background (ATSI), Speaking English (NESB or HOME), Living in Australia (INOZ) and Transience (TRANS). However, the variable TRANS is available for examination only in models that include all the students who could be matched regardless of whether they had remained in the same school or changed schools between Grade 3 and Grade 5 levels.

In HLM analyses, each student background variable included in $\boldsymbol{X}_{hitj}$ is grand-mean centred, therefore, the estimates of the intercepts, $\boldsymbol{\pi}_{0tj}$, are the intake-adjusted school means. Hence, the intercepts describe how well a student with sample-average

background characteristics can be expected to score in a given school (Willms and Raudenbush, 1989; Kreft, 1995; Kreft et al., 1995; Raudenbush and Bryk, 1997).

Level-2 model

The meso-level of the model for the intercepts regresses the intake-adjusted performance, $\boldsymbol{\pi}_{0tj}$, on $\mathbf{OCC}_{tj}$, the t^{th}-testing occasion for each school.

$$\boldsymbol{\pi}_{0tj} = \boldsymbol{\beta}_{00j} + \boldsymbol{\beta}_{01j}\mathbf{OCC}_{tj} + \mathbf{r}_{0tj} \qquad \text{Equation 9.2}$$

The time trend variable OCC is group-mean centred, therefore:

$\boldsymbol{\beta}_{00j}$ is the mean effectiveness of school *j* during the period of the study (see Kreft, 1995; Kreft et al., 1995),

$\boldsymbol{\beta}_{01j}$ is the difference in school *j* trend in achievement relative to the overall trend,

$\mathbf{r}_{0tj}$ is a random year-to-year fluctuation in a school's intake-adjusted levels of performance.

The meso-level of the Type A effects model, therefore, decomposes intake-adjusted levels of performance ($\boldsymbol{\pi}_{0tj}$) into a stable component ($\boldsymbol{\beta}_{00j}$) and a component that varies across occasions ($\boldsymbol{\beta}_{01j}\mathbf{OCC}_{tj} + \mathbf{r}_{0tj}$).

In addition, at this level of the model each component that is associated with the student background characteristics, ($\boldsymbol{\pi}_{htj}$) is viewed as an outcome varying randomly around some school mean ($\boldsymbol{\beta}_{h0j}$), that is:

$$\boldsymbol{\pi}_{1tj} = \boldsymbol{\beta}_{10j} + \mathbf{r}_{1tj}$$

$$\boldsymbol{\pi}_{2tj} = \boldsymbol{\beta}_{20j} + \mathbf{r}_{2tj}$$

.

.

.

$$\boldsymbol{\pi}_{htj} = \boldsymbol{\beta}_{h0j} + \mathbf{r}_{htj} \qquad \text{Equation 9.3}$$

Level-3 model

At the macro-level, the mean performance level $\boldsymbol{\beta}_{00j}$ of each school is viewed as varying randomly around a grand mean:

$$\boldsymbol{\beta}_{00j} = \boldsymbol{\gamma}_{000} + \mathbf{u}_{00j} \qquad \text{Equation 9.4}$$

where:

$\boldsymbol{\beta}_{00j}$ is the mean effectiveness of school *j*,

$\boldsymbol{\gamma}_{000}$ is the grand mean

$\mathbf{u}_{00j}$ is a random 'school effect', that is, the deviation of mean effectiveness of the school from the grand mean.

Similarly, the variation between schools in their trend component, $\boldsymbol{\beta}_{01j}$, is represented as an overall mean, $\boldsymbol{\gamma}_{010}$, and a random component, $\mathbf{u}_{01j}$:

$$\boldsymbol{\beta}_{01j} = \boldsymbol{\gamma}_{010} + \mathbf{u}_{01j} \qquad \text{Equation 9.5}$$

For parsimony, the effects associated with the student background characteristics, $\boldsymbol{\beta}_{h0j}$, are specified as fixed, that is, the errors terms ($\mathbf{u}_{h0j}$) of the student background components are deleted from the model:

$\beta_{10j} = \gamma_{100}$

$\beta_{20j} = \gamma_{200}$

.

.

.

$\beta_{h0j} = \gamma_{h00}$ Equation 9.6

However, in the actual analyses, the effects associated with the student background characteristics are only specified as fixed if they do not vary significantly across the schools or across the occasions or if their reliability estimates fall below the 0.05 value recommended by Bryk and Raudenbush (1992).

In two steps, Equations 9.1, 9.2, 9.4 and 9.5 can be combined into a single equation to yield a model, which describes the linear relationship of the components involved. The first step involves the substitution for π_{0tj} (from Equation 9.2) in Equation 9.1 to give the following equation:

$$\mathbf{Y}_{itj} = \beta_{00j} + \beta_{01j}\mathbf{OCC}_{tj} + \mathbf{r}_{0tj} + \pi_{htj}\boldsymbol{X}_{hitj} + \mathbf{e}_{itj} \quad \text{Equation 9.7}$$

The second step involves the substitution for β_{00j} and β_{01j} (from Equations 9.4 and 9.5, respectively) in Equation 9.7 to provide the following equation:

$$\mathbf{Y}_{itj} = \{\gamma_{000} + \mathbf{u}_{00j}\} + \{(\gamma_{010} + \mathbf{u}_{01j})\mathbf{OCC}_{tj} + \mathbf{r}_{0tj}\} + \pi_{htj}\boldsymbol{X}_{hitj} + \mathbf{e}_{itj}$$

simplifying gives

$$\mathbf{Y}_{itj} = \{\gamma_{000} + \mathbf{u}_{00j}\} + \{\gamma_{010}\,\mathbf{OCC}_{tj} + \mathbf{u}_{01j}\mathbf{OCC}_{tj} + \mathbf{r}_{0tj}\} + \pi_{htj}\boldsymbol{X}_{hitj} + \mathbf{e}_{itj}$$

or

$$\mathbf{Y}_{itj} = [\gamma_{000}] + [\gamma_{010}\,\mathbf{OCC}_{tj}] + [\pi_{htj}\boldsymbol{X}_{hitj}] + [\mathbf{u}_{00j}] + [\mathbf{u}_{01j}\mathbf{OCC}_{tj} + \mathbf{r}_{0tj}] + [\mathbf{e}_{itj}]$$

that is,

$$\begin{aligned}
\mathbf{Y}_{itj} &= [\gamma_{000}] && \text{(grand mean)}\\
&+ [\gamma_{010}\,\mathbf{OCC}_{tj}] && \text{(main effect of occasion)}\\
&+ [\pi_{htj}\boldsymbol{X}_{hitj}] && \text{(control for student intake)}\\
&+ [\mathbf{u}_{00j}] && \text{(stable component of school effect)}\\
&+ [\mathbf{u}_{01j}\mathbf{OCC}_{tj} + \mathbf{r}_{0tj}] && \text{(change component of school effect)}\\
&+ [\mathbf{e}_{itj}] && \text{(student-level random error)}
\end{aligned}$$

Equation 9.8

The component $\mathbf{u}_{00j}$ in the above model (Equation 9.8) for Type A effects represents the increment to student achievement attributable to school *j* and is constant across occasions after controlling for the effects of student intake. This component includes the effects that can be attributed to both school context and to school policy and, therefore, it is the stable Type A effect (Raudenbush and Willms, 1995).

In Equation 9.8, the change component of school effect has two terms: (a) $\mathbf{u}_{01j}\mathbf{OCC}_{tj}$, which represents a 'school-by-occasion' interaction effect, and (b) $\mathbf{r}_{0tj}$, which represents a random year-to-year fluctuation in a school's intake-adjusted levels of performance. Of importance here is the value of the school-by-occasion interaction effect ($\mathbf{u}_{01j}\mathbf{OCC}_{tj}$) because it shows how a school has performed over time. A positive value of $\mathbf{u}_{01j}\mathbf{OCC}_{tj}$ indicates that achievement changed more than expected in school *j*

while a negative value indicates that achievement changed less than expected in school *j* (Willms and Raudenbush, 1989). The value of the term $\mathbf{u}_{01j}\mathbf{OCC}_{tj}$, therefore, shows the change (improvement or deterioration) that has occurred in a school's Type A effect over the study period.

Specification of Type B effects model

For the current study, modelling for Type B effects involves (a) student background characteristics (Type A effect), (b) student-related school-level variables, that is, school context, and (c) averaged school context over the study period.

Figures 9.1 and 9.2 show the proposed longitudinal models for estimation of Type B using all the students who could be matched (transience, N=37,832) and the students who remained in the same school (non-transience, N=32,741), respectively. The variables examined for inclusion in these models are described in Chapter 5.

At the micro-level, the models shown in Figures 9.1 and 9.2 are the same as the corresponding models for the estimation of Type A effects described above. However, unlike the Type A effects, predictors are included in the models for the estimation of Type B effects at the meso-level and macro-level as shown in the two diagrams. But it should be borne in mind that Type A and Type B effects models have the linear trend variable (OCC) included at the meso-level.

The three-level model for the estimation of Type B effects based on a longitudinal structure is described next in equation format.

Level-1 model

At the micro-level, the model for Type B effects is the same as the model for Type A effects, that is:

$$\mathbf{Y}_{itj} = \boldsymbol{\pi}_{0tj} + \boldsymbol{\pi}_{htj}\boldsymbol{X}_{hitj} + \mathbf{e}_{itj} \qquad \text{Equation 9.9}$$

The components $\mathbf{Y}_{itj}$, $\boldsymbol{\pi}_{0tj}$, $\boldsymbol{\pi}_{htj}\boldsymbol{X}_{hitj}$, and $\mathbf{e}_{itj}$ in Equation 9.9 have the same meaning they carried in Equation 9.1 given above for the Type A effects model.

Level-2 model

At the meso-level for the Type B effects model, the intake-adjusted performance, $\boldsymbol{\pi}_{0tj}$, is regressed on school-context ($\bar{\boldsymbol{X}}_{gtj}$) variables that change between occasions, and on $\mathbf{OCC}_{tj}$, the t^{th}-testing occasion for each school.

$$\boldsymbol{\pi}_{0tj} = \boldsymbol{\beta}_{00j} + \boldsymbol{\beta}_{01j}\mathbf{OCC}_{tj} + \boldsymbol{\beta}_{0gj}\bar{\boldsymbol{X}}_{gtj} + \mathbf{r}_{0tj} \qquad \text{Equation 9.10}$$

where:

$\boldsymbol{\beta}_{0gj}$, are the slopes associated with the changing school context; and

$\boldsymbol{\pi}_{0tj}$, $\boldsymbol{\beta}_{00j}$ and $\mathbf{r}_{0tj}$ carry the same meaning as described above for the model for Type A effects (Equation 9.2).

For parsimony, $\boldsymbol{\beta}_{0gj}\bar{\boldsymbol{X}}_{gtj}$ in Equation 9.10 represents the control for several relevant school-level variables ($\boldsymbol{\beta}_{02j}\bar{\boldsymbol{X}}_{gtj} + \boldsymbol{\beta}_{03j}\bar{\boldsymbol{X}}_{gtj} + \ldots + \boldsymbol{\beta}_{0gj}\bar{\boldsymbol{X}}_{gtj}$) that describe the school context at each testing occasion. There are g = 2, 3, 4, . . . , G school-level variables which describe the changing school context. The measures of the school context are the student-related school-level variables. At each testing occasion, the variables that describe the changing school context are formed by either aggregating student characteristic variables or from the school information data set obtained from the DETE in South Australia. Hence, for the current study $\bar{\boldsymbol{X}}_{gtj}$ represents a combination of any of the following student-related school-level variables: Y3NSCO_1 or

Y3LSCO_1, SEX_1, AGE_1, ATSI_1, NESB_1 or HOME_1, INOZ_1, TRANS_1 or MOBILITY, PSCARD, and ABSENT. The amount of between-school variance left after controlling for the student-related school-level variables can be attributed directly to the schools as such rather than the students who attend them. Hence, the student free school-level variables, such as School Size (SSIZE) and School Location (METRO or GPODIST), are excluded in order to estimate how much of the variance is taken up by the characteristics of the student population in the school (Harker and Nash, 1996).

In HLM analyses, the time trend variable (OCC) and each school context variable ($\bar{X}_{gtj}$) are group-mean centred, therefore, the intercept, $\beta_{00j,}$ is the mean effectiveness of school *j* during the period of the study (Kreft, 1995; Kreft et al., 1995).

Similar to the Type A effects model, the meso-level of the model for Type B effects decomposes intake-adjusted levels of performance (π_{0tj}) into a stable component (β_{00j}) and a component that varies across occasions ($\beta_{01j}\mathbf{OCC}_{tj} + \beta_{0gj}\bar{X}_{gtj} + \mathbf{r}_{0tj}$). Likewise, this level of the Type B effects model views each component associated with the student background characteristics, (π_{htj}) as an outcome varying randomly around some school mean (β_{h0j}), that is:

$$\pi_{1tj} = \beta_{10j} + \mathbf{r}_{1tj}$$

$$\pi_{2tj} = \beta_{20j} + \mathbf{r}_{2tj}$$

.

.

.

$$\pi_{htj} = \beta_{h0j} + \mathbf{r}_{htj} \qquad \text{Equation 9.11}$$

Equation 9.11 allows the examination of possible interaction effects between the student-level variables and the school-level variables.

Level-3 model

At the macro-level, the mean performance level β_{00j} of each school is regressed on the average components of school context, $\acute{x}_{00fj}$:

$$\beta_{00j} = \gamma_{000} + \gamma_{00f}\acute{x}_{00fj} + \mathbf{u}_{00j} \qquad \text{Equation 9.12}$$

where:

γ_{00f} are the slopes associated with the average school context over the study period; and

β_{00j}, γ_{000}, and $\mathbf{u}_{00j}$ carry the same meaning as described above for the Type A effects model (Equation 9.4).

For parsimony, the component $\gamma_{00f}\acute{x}_{00fj}$ is used to represent control for several school context variables ($\gamma_{001}\acute{x}_{001j} + \gamma_{002}\acute{x}_{002j} + \ldots + \gamma_{00f}\acute{x}_{00fj}$) that are each formed by averaging the relevant school context variables ($\bar{X}_{gtj}$) over the four testing occasions. There are $f = 1, 2, \ldots, F$ school-level variables which describe the average school context. Hence, $\acute{x}_{00j}$ represents a combination of any of the following variables, which describe the average school context over the duration of the study: Y3NSCO_2 or Y3LSCO_2, SEX_2, AGE_2, ATSI_2, NESB_2 or HOME_2, INOZ_2, TRANS_2 or MOBILI_2, PSCARD_2, and ABSENT_2. Again, the school based variables (for example, School Location), are excluded in order to estimate how much of the

variance is taken up by the characteristics of the pupil population in the school (Harker and Nash, 1996).

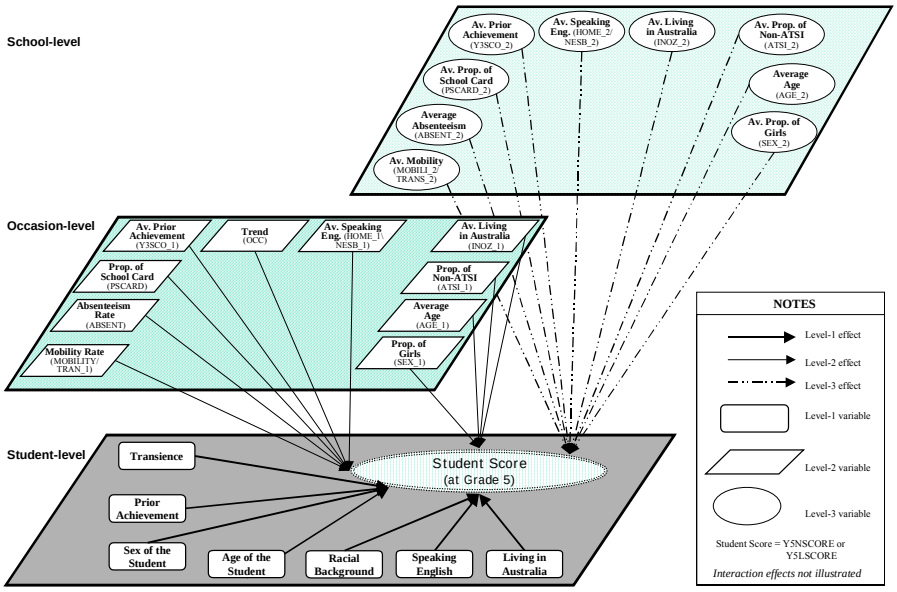


Figure 9.1 Type B effects model using the transience data set

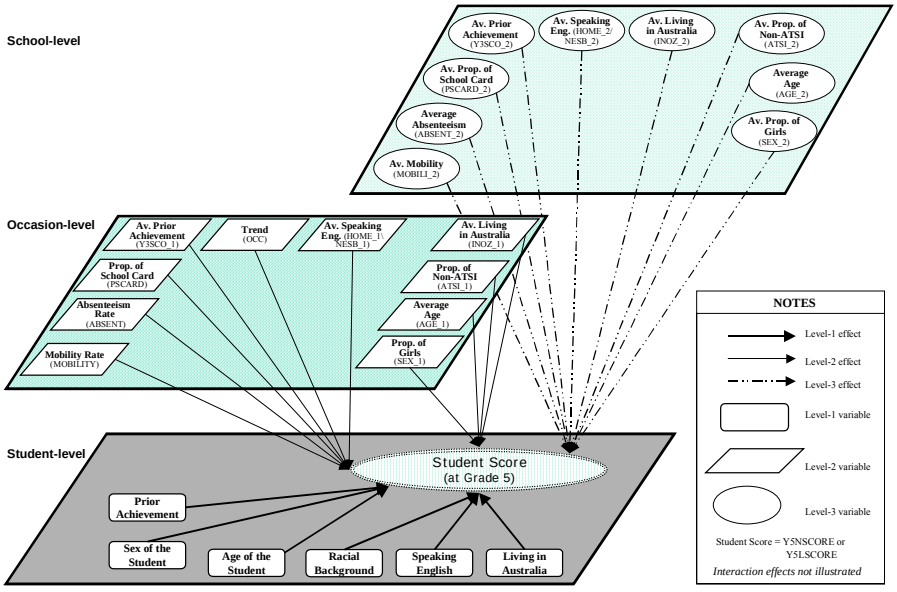


Figure 9.2 Type B effects model using the non-transience data set

In addition, at the macro-level, the variation between schools in their trend component, β_{01j}, is regressed on the average school context, $\acute{x}_{01fj}$:

$$\beta_{01j} = \gamma_{010} + \gamma_{01f}\acute{x}_{01fj} + u_{01j} \qquad \text{Equation 9.13}$$

Thus, Equation 9.13 allows the examination of how much of the instability can be explained by the schools' context.

Again for parsimony, the effects associated with the student background characteristics, β_{h0j}, and the effects associated with changing school context, β_{0gj}, are specified as fixed, that is, the errors terms (u) of the student background and school context components are deleted from the model. In the actual analyses, these effects are only specified as fixed if they do not vary significantly across the schools and/or across the occasions or if the values of reliability estimates of the variable are small.

Using the same procedure followed to develop a single linear equation for the estimation of Type A effects, Equations 9.9, 9.10, 9.12 and 9.13 can be combined into a single equation as follows:

1). substitution for π_{0tj} (from Equation 9.10) in Equation 9.9

$$Y_{itj} = \beta_{00j} + \beta_{01j}OCC_{tj} + \beta_{0gj}\bar{X}_{gtj} + r_{0tj} + \pi_{htj}X_{hitj} + e_{itj} \qquad \text{Equation 9.14}$$

2). substitution for β_{00j} and β_{01j} (from Equations 9.12 and 9.13, respectively) in Equation 9.6

$$Y_{itj} = \{\gamma_{000} + \gamma_{00f}\acute{x}_{00fj} + u_{00j}\} + \{\gamma_{010} + \gamma_{01f}\acute{x}_{01fj} + u_{01j}\}OCC_{tj} + \beta_{0gj}\bar{X}_{gtj} + r_{0tj} + \pi_{htj}X_{hitj} + e_{itj}$$

simplifying gives

$$Y_{itj} = \{\gamma_{000} + \gamma_{00f}\acute{x}_{00fj} + u_{00j}\} + \{\gamma_{010}OCC_{tj} + \gamma_{01f}\acute{x}_{01fj}OCC_{tj} + u_{01j}OCC_{tj}\} + \beta_{0gj}\bar{X}_{gtj} + r_{0tj} + \pi_{htj}X_{hitj} + e_{itj}$$

or

$$Y_{itj} = [\gamma_{000}] + [\gamma_{010}OCC_{tj}] + [\pi_{htj}X_{hitj}] + [\gamma_{00f}\acute{x}_{00fj} + u_{00j}] + [\gamma_{01f}\acute{x}_{01fj}OCC_{tj} + u_{01j}OCC_{tj} + \beta_{0gj}\bar{X}_{gtj} + r_{0tj}] + e_{itj}$$

that is,

$$\begin{aligned} Y_{itj} &= [\gamma_{000}] && \text{(grand mean)} \\ &+ [\gamma_{010}OCC_{tj}] && \text{(main effect of occasion)} \\ &+ [\pi_{htj}X_{hitj}] && \text{(control for student intake)} \\ &+ [\gamma_{00f}\acute{x}_{00fj} + u_{00j}] && \text{(stable component of school effect)} \\ &+ [\gamma_{01f}\acute{x}_{01fj}OCC_{tj} + u_{01j}OCC_{tj} + \beta_{0gj}\bar{X}_{gtj} + r_{0tj}] && \text{(change component of school effect)} \\ &+ [e_{itj}] && \text{(student-level random error)} \end{aligned}$$

Equation 9.15

In Equation 9.15, the stable component of school effect now has two terms: $\gamma_{00f}\acute{x}_{00fj}$ and u_{00j}. The term $\gamma_{00f}\acute{x}_{00fj}$ represents the control for average school context over the duration of the study while the residual term u_{00j} now represents the increment to student achievement attributable to school *j* after controlling for the effects of student intake and the effects of the average school context. Hence, this residual term (u_{00j}) includes mostly the effects attributed to school characteristics and school policy and, therefore, it is the stable Type B effect (Raudenbush and Willms, 1995).

In the above model, the change component of school effect has four terms: (a) $\gamma_{01f}\acute{x}_{01fj}OCC_{tj}$, a 'context-by-occasion' interaction effect, (b) $u_{01j}OCC_{tj}$, a 'school-by-

occasion' interaction effect, (c) $\beta_{0gj}\bar{X}_{gtj}$, control for changing school context and (d) $\mathbf{r}_{0tj}$, a random year-to-year fluctuation in a school's intake-adjusted levels of performance. Here again the main interest is concern with the value of the 'school-by-occasion' interaction effect ($\mathbf{u}_{01j}\mathbf{OCC}_{tj}$), because it shows the change that has occurred in a school's Type B effect over the study period. That is, it represents the systematic change in the performance of the school after allowance has been made for student characteristics and school context.

Estimation of Type A effects

The model for Type A effects specified above (Equation 9.8) is estimated using two data sets. The estimation is first carried out using all students who have two data points irrespective of whether or not the students had changed schools and then the estimation is repeated using only those students who remained in the same school between the two grade levels. For each of the two data sets used, two separate models are estimated, one for numeracy and the other for literacy.

Harker and Nash's (1996) approach as well as the approach of Thomas et al. (1997) are to keep a student-level variable in the analysis even if the variable makes no significant contribution overall, on the grounds that it may have a substantial effect on the estimation of coefficients and residuals for specific schools. However, because of the large number of cases (over 32,000) involved in the current analyses, significance of the student-level variables is considered essential in this study for inclusion of the variable into the analysis. Consequently, the step-up approach is followed to examine which of the student-level variables have a significant ($p<0.05$) influence on achievement in numeracy and literacy using each of the two data sets. Any non-significant student-level variables are excluded from the analysis. In addition, the regression coefficients of the student-level variables that do not vary significantly at Levels 2 and 3 (or that have low [≤0.05] reliability estimates) are modelled as fixed at those levels (Willms and Raudenbush, 1989; Harker and Nash, 1996).

Using the step-up approach, Type A effects are estimated by controlling only for student characteristics, leaving school context and policy unspecified. This procedure enables the school effects to be estimated by the subtraction method (see Chapter 4 and also Raudenbush and Willms 1995).

An example of the models used to estimate the Type A effects is presented below in equation form.

Level-1 model

$$[\mathbf{Y5NSCORE}]_{itj} = \pi_{0tj} + \pi_{1tj}\mathbf{SEX}_{itj} + \pi_{2tj}\mathbf{TRANS}_{itj} + \pi_{3tj}\mathbf{AGE}_{itj} + \pi_{4tj}\mathbf{ATSI}_{itj} + \pi_{5tj}\mathbf{INOZ}_{itj} + \pi_{6tj}\mathbf{Y3NSCORE}_{itj} + \mathbf{e}_{itj}$$

Level-2 model

$$\pi_{0tj} = \beta_{00j} + \beta_{01j}(\mathbf{OCC})_{tj} + \mathbf{r}_{0tj}$$
$$\pi_{1tj} = \beta_{10j}$$
$$\pi_{2tj} = \beta_{20j}$$
$$\pi_{3tj} = \beta_{30j}$$
$$\pi_{4tj} = \beta_{40j}$$
$$\pi_{5tj} = \beta_{50j}$$
$$\pi_{6tj} = \beta_{60j} + \mathbf{r}_{60tj}$$

Level-3 model

$$\beta_{00j} = \gamma_{000} + u_{00j}$$
$$\beta_{01j} = \gamma_{010} + u_{01j}$$
$$\beta_{10j} = \gamma_{100} + u_{10j}$$
$$\beta_{20j} = \gamma_{200} + u_{20j}$$
$$\beta_{30j} = \gamma_{300}$$
$$\beta_{40j} = \gamma_{400}$$
$$\beta_{50j} = \gamma_{500}$$
$$\beta_{60j} = \gamma_{600}$$

Equation 9.16

Estimation of Type B effects

The model for Type B effects specified above (Equation 9.15) is estimated for numeracy and literacy using the two data sets. Technically, the estimation of Type B effects involves undertaking further HLM runs on the models presented above for estimation of Type A effects (For example, Equation 9.16) to control for the significant contextual Levels 2 and 3 variables using the step-up strategy. At this stage, the exploratory analysis sub-routine is employed for examining the inclusion of potentially significant Levels 2 and 3 predictors in successive HLM runs. The student-free school-level variables, such as School Size and School Location, are excluded in order to estimate how much of the variance is taken up by the characteristics of the pupil population in the school (Harker and Nash, 1996).

In addition, at this stage, the existence of any significant ($p<0.05$) cross-level interaction effects is examined for all variables that are modelled as varying at Level-2 or/and at Level-3. A variable might show significant variation in Type A effects model but no significant variation after adding the school context variables, and therefore, the error term of such a variable is included in the Type A effects model and deleted in the Type B effects model.

As an example, the final model used for the estimation of the Type B effects for numeracy using all students who could be matched is presented below (Equation 9.17).

Level-1 model

$$[\mathbf{Y5NSCORE}]_{itj} = \pi_{0tj} + \pi_{1tj}\mathbf{SEX}_{itj} + \pi_{2tj}\mathbf{TRANS}_{itj} + \pi_{3tj}\mathbf{AGE}_{itj} + \pi_{4tj}\mathbf{ATSI}_{itj} + \pi_{5tj}\mathbf{INOZ}_{itj} + \pi_{6tj}\mathbf{Y3NSCORE}_{itj} + e_{itj}$$

Level-2 model

$$\pi_{0tj} = \beta_{00j} + \beta_{01j}\mathbf{OCC}_{tj} + \beta_{02j}\mathbf{AGE\_1}_{tj} + \beta_{03j}\mathbf{HOME\_1}_{tj} + r_{0tj}$$
$$\pi_{1tj} = \beta_{10j}$$
$$\pi_{2tj} = \beta_{20j}$$
$$\pi_{3tj} = \beta_{30j}$$
$$\pi_{4tj} = \beta_{40j}$$
$$\pi_{5tj} = \beta_{50j}$$
$$\pi_{6tj} = \beta_{60j} + \beta_{61j}\mathbf{INOZ\_1}_{tj} + r_{60tj}$$

Level-3 model

$$\beta_{00j} = \gamma_{000} + \mathbf{u}_{00j}$$
$$\beta_{01j} = \gamma_{010} + \mathbf{u}_{01j}$$
$$\beta_{02j} = \gamma_{020}$$
$$\beta_{03j} = \gamma_{030}$$
$$\beta_{10j} = \gamma_{100} + \gamma_{101}\mathbf{Y3NSCO\_2}_j + \mathbf{u}_{10j}$$
$$\beta_{20j} = \gamma_{200} + \gamma_{201}\mathbf{AGE\_2}_j + \gamma_{202}\mathbf{Y3NSCO\_2}_j + \mathbf{u}_{20j}$$
$$\beta_{30j} = \gamma_{300}$$
$$\beta_{40j} = \gamma_{400}$$
$$\beta_{50j} = \gamma_{500}$$
$$\beta_{60j} = \gamma_{600}$$
$$\beta_{61j} = \gamma_{610}$$

Equation 9.17

By definition, at Level-1, the models for the estimation of Type A effects (for example, Equation 9.16) are identical to the corresponding models for the estimation of Type B effects (for example, Equation 9.17). Predictably, at Level-1, the models for estimation of school effects for numeracy and literacy obtained in this chapter are exactly the same as the two- and three-level models examined in earlier chapters. Regardless of the data set used, five student-level variables have a significant influence on both numeracy and literacy. These five variables are namely Sex of the Student (SEX), Age of the Student (AGE), Racial Background (ATSI), Living in Australia (INOZ) and Prior Achievement (Y3NSCORE for numeracy or Y3LSCORE literacy). In addition, the variable Speaking English at Home (HOME) has a significant influence on literacy but not on numeracy, and where examined, the variable Transience (TRANS) has a significant influence on both numeracy and literacy.

Except for the effects of Prior Achievement (Y3NSCORE or Y3LSCORE), the effects of all the other student-level variables are modelled as fixed across the four occasions (that is, at Level-2). The effects of a variable are modelled as fixed if they do not vary significantly across the four occasions, or if the variable's reliability estimate is small (≤0.05), or the program has problems converging after the variable is added into the equation (Bryk and Raudenbush, 1992; Hox, 1995).

For the literacy models, it should be noted that the error terms for variable Racial Background (ATSI) are deleted in Type B effects models but not in Type A effects models because the variable does not show significant variation after the school context variables are added. For the same reason, the variable AGE is specified as fixed in the Type B effects model but varying in the Type A effects model for numeracy when considering students who were matched in the same school, that is, the non-transience data set.

Results

The results of the above HLM analyses provide reliability estimates at Levels 1 and 2 of the model for each variable with random effects at those levels. The results also provide estimations of the fixed effects for each variable in the equation, estimations of the variance components and the deviance statistics of the models. These results are discussed in separate sub-sections below.

Reliability estimates

Table 9.1 displays the school-level reliability estimates of the stable and the change components of the simplest longitudinal (null) models, Type A (second panel of Table 9.1) and Type B (third panel of Table 9.1) effects models. The reliability estimates of all the variables with random effects at the second and the third levels of these models are given in Appendix 14.3.

The reliability estimates from the simplest longitudinal model (results in the first panel of Table 9.1) show that the stable components are estimated far more reliably than change effects. For example, for numeracy and using the transience data set (that is, using all the students who could be matched), the reliability estimate for the stable component is 0.859, and for the change component, it is 0.092. For this example, 85.9 per cent and 9.2 per cent of the variance among estimates of intercepts and occasion slopes respectively, can be considered true parameter variance; the remaining 14.1 per cent and 90.8 per cent respectively, are random fluctuations that could be associated with measurement and sampling error. Hence, this data set contains substantial "signal" (Bryk and Raudenbush, 1992; p. 137) for detecting differences between schools in their stable school effects with less signal for detecting differences in change school effects. However, prediction of differences in change effects is still warranted because results of preliminary analyses from the unconditional model reveal that there are statistically significant differences between schools on occasion slopes as well as average performances (Raudenbush, 1995). Furthermore, (Bryk and Raudenbush (1992) reported that it is generally possible to undertake HLM estimations with reliabilities as low as 0.05.

Table 9.1 School-level reliability estimates from Type A and Type B effects models

			Numeracy		Literacy	
Model			Tran$^{\propto}$	Non-Tran$^{\beta}$	Tran$^{\propto}$	Non-Tran$^{\beta}$
	Effect	***Random level-2 coefficient***				
Null	Stable	**INTRCPT1**/ INTRCPT2,	0.859	0.836	0.844	0.825
	Change	**INTRCPT1**/ OCC,	0.092	0.053	0.106	0.116
Type A	Stable	**INTRCPT1**/ INTRCPT2,	0.635	0.602	0.533	0.459
	Change	**INTRCPT1**/ OCC	0.178	0.173	0.193	0.170
Type B	Stable	**INTRCPT1**/ INTRCPT2,	0.435	0.445	0.403	0.372
	Change	**INTRCPT1**/ OCC	0.176	0.171	0.162	0.145

Notes: $\propto$ - Using all the students matched (Schools = 482).
β - Using only those students matched in the same school (Schools = 479).

Compared to the simplest longitudinal model, the results in Table 9.1 show that the inclusion of predictors into the models lowers the reliability estimates of the stable components but improves the reliability estimates of the change components. Nevertheless, the results indicate that the reliability estimates of the stable components are substantially higher than the reliability estimates of the change components with or without inclusion of predictors into the models. That is, the results show that in either the Type A or the Type B models, the stable effects are still estimated far more reliably than change effects. For example, for numeracy and using the transience data set, the reliability estimate for Type A effect is 0.635, and for change effect, it is

0.178. These results are consistent whether the transience data set is considered or whether the non-transience data set is considered, and consistent across the two subject areas included in the BSTP.

For the same subject area, the results in Table 9.1 show that the reliability estimates of the stable Type A or Type B and change effects observed using the transience data set follow closely those obtained using the non-transience data set. However, the reliability estimates of the Type A effect are substantially higher than for the Type B effect. For example, for numeracy and using all students who could be matched, the reliability estimates for the Type A effect is 0.635, while the corresponding estimate for the Type B effect is noticeably lower (0.435). Hence, as it would be expected, Type A effects are estimated more reliably than are Type B effects (Willms and Raudenbush, 1989).

In general, the results in second and third panels of Table 9.1 also show that the reliability estimates for numeracy are slightly higher than for literacy especially for the stable school effects. For example, for numeracy and using the non-transience data set, the reliability estimates for Type A effect is 0.602, while the corresponding estimate for literacy is 0.459. Hence, it appears that the stable school effects for numeracy are estimated more reliably than for literacy.

Finally, it should also be noted that the reliability estimates obtained for change effect using the Type A effect model (results in the second panel of Table 9.1) follow closely the estimates obtained using the Type B effect models (results in the third panel of Table 9.1). For example, for numeracy and the transience data set, the reliability estimate for change effect is 0.178 for the Type A effects model, while the estimate is 0.176 for the Type B effects model. Hence, it appears that the adjustment for school context does not substantially affect the reliability of the change effect.

Deviance statistics

Table 9.2 presents results of deviance statistics and the chi-square tests carried out to compare model fit for both types of school effects. In Table 9.2, the fit of the Type A effects model is compared to the fit of the simplest longitudinal model (null), and the fit of the Type B effects model is compared to the fit of the Type A effects model.

The chi-square tests presented in Table 9.2 indicate better fit of the Type A effects models compared to the simplest longitudinal models and better fit of the Type B effects models compared to the Type A effects models for both outcome measures regardless of the data set used. Therefore, the inclusion of the predictors at the three levels of hierarchy significantly improves the overall fit of the models.

Fixed effects

The estimations of the fixed effects for the models for Type A and Type B effects are presented in Tables 9.3 and 9.4, respectively. Both the standardized as well as the metric regression coefficients of the variables in the final models for Type A and Type B effects are presented in the two tables.

The results shown in Tables 9.3 and 9.4 and the results of the analyses presented in the preceding chapters indicate the same relationships regarding student-level factors influencing achievement in numeracy and literacy among Grade 5 students in South Australia. Prior Achievement is positively related to achievement at Grade 5. Boys outperformed girls in numeracy but girls outperformed them in literacy. For numeracy as well as literacy, younger pupils outperformed their older counterparts, non-ATSI students outperformed ATSI students, new students to Australia outperformed

students who were born in Australia, and non-transience students outperformed transience students. In addition, students from an English-speaking background outperformed students from a non-English-speaking background in literacy but not in numeracy.

Table 9.2 Comparison of model fit using chi-square tests

		Deviance Statistic	Number of Parameters	Chi-square Statistic	Degrees of Freedom	P-value
Numeracy						
Tran$^{\propto}$	Null^{c}	109,262.10	7			
	Type A	88,113.21	22	21,148.89	15	0.00
	Type B	87,773.15	31	340.06	9	0.00
Non-Tran$^{\beta}$	Null^{c}	94,220.61	7			
	Type A	75,351.48	21	18,869.14	14	0.00
	Type B	75,168.28	19	183.20	2	0.00
Literacy						
Tran$^{\propto}$	Null^{c}	109,183.51	7			
	Type A	79,918.32	28	29,265.19	21	0.00
	Type B	79,776.80	27	141.52	1	0.00
Non-Tran$^{\beta}$	Null^{c}	93,809.04	7			
	Type A	67,944.12	27	25,864.92	20	0.00
	Type B	67,861.54	24	82.59	3	0.00

Notes: $\propto$ - Using all the students matched (Schools = 482).
β - Using only those students matched in the same school (Schools = 479).
c - Simplest longitudinal model

In the HLM analyses described in this chapter, the predictor OCC is group-mean centred. Therefore, the regression coefficient B01 (Tables 9.3 and 9.4) represents the average change in achievement across all schools over the study period (Kreft, 1995; Kreft et al., 1995). Hence, the negative coefficients and significant t-ratios values for OCC in Tables 9.3 and 9.4 indicate that there are general declines in the performance of the schools between occasions. That is, the performance of schools on the earlier occasions was estimated to be higher than their performance on the later occasions. This finding is consistent across the two types of school effects, across the two subject areas and across the two data sets. Arguably, these results provide a better picture of the performance of the schools over time than the results of the two-level and three-level analyses reported in the Chapters 7 and 8 respectively. This is because the longitudinal structure employed in this chapter has allowed schools to be linked over time and the number of units at the school-level is large (482 and 479 for the transience and non-transience data sets respectively) for significance testing.

At the meso-level, the results for Type B effects (Table 9.4) indicate that no measure of changing school context has consistent and significant effects on the two outcome measures. Indeed, the results show that no school context variable (except Average Age of the Students [AGE_1] and Average Speaking English at Home [HOME_1] for numeracy when using the transience data set) has significant ($p<0.05$) effects at the meso-level.

Table 9.3 Final estimation of fixed effects from Type A effects models

		Transience Data Set (Schools = 482)					**Non-Transience Data Set** (Schools = 479)				
		Std'zed	Metric	SE	T-ratio	P-value	Std'zed	Metric	SE	T-ratio	P-value
Numeracy											
INTRCPT1,											
INTRCPT2	B00	1.36	1.36	0.01	113.52	0.00 *kj*	1.41	1.41	0.01	117.80	0.00 *kj*
OCC	B01	-0.03	-0.02	0.01	-3.52	0.00 *j*	-0.03	-0.02	0.01	-3.18	0.00 *j*
SEX	B10	-0.04	-0.09	0.01	-9.89	0.00 *j*	-0.05	-0.10	0.01	-10.62	0.00 *j*
AGE	B20	-0.07	-0.20	0.01	-16.19	0.00	-0.07	-0.19	0.01	-14.79	0.00 *j*
ATSI	B30	0.04	0.23	0.03	8.75	0.00	0.04	0.24	0.03	7.97	0.00
INOZ	B40	-0.01	-0.04	0.01	-2.91	0.00	-0.01	-0.04	0.01	-2.92	0.00
TRANS	B50	-0.04	-0.10	0.01	-7.73	0.00 *j*					
Y3NSCORE	B60	0.69	0.56	0.01	118.29	0.00 *k*	0.71	0.57	0.01	117.65	0.00 *k*
Literacy											
INTRCPT1,											
INTRCPT2	B00	1.48	1.48	0.01	149.39	0.00 *kj*	1.53	1.53	0.01	158.40	0.00 *kj*
OCC	B01	-0.11	-0.10	0.01	-15.70	0.00 *j*	-0.11	-0.10	0.01	-15.20	0.00 *j*
SEX	B10	0.02	0.04	0.01	5.36	0.00 *j*	0.02	0.03	0.01	3.92	0.00 *j*
AGE	B20	-0.06	-0.18	0.01	-16.50	0.00	-0.06	-0.17	0.01	-15.40	0.00 *j*
ATSI	B30	0.03	0.18	0.03	6.68	0.00 *j*	0.03	0.16	0.03	5.50	0.00
HOME	B40	0.02	0.03	0.01	3.80	0.00	0.01	0.03	0.01	3.17	0.00
INOZ	B50	-0.02	-0.05	0.01	-4.47	0.00	-0.02	-0.06	0.01	-4.55	0.00
TRANS	B60	-0.02	-0.06	0.01	-4.70	0.00					
Y3LSCORE	B70	0.78	0.61	0.00	145.32	0.00 *kj*	0.78	0.63	0.00	143.88	0.00 *kj*

Notes: Shade - The variable TRANS is not available for examination in this model.
Std'zed - Regression coefficient obtained using standardized variables.
Metric - Regression coefficient obtained using unstandardized variables.
- Standard errors (SE), t-ratios and p-values presented are those obtained using unstandardized variables.
k - Residual parameter of this coefficient is left to vary at the occasion-level.
j - Residual parameter of this coefficient is left to vary at the school-level.

Table 9.4 Final estimation of fixed effects from Type B effects models

1) Numeracy

			Transience Data Set (Schools = 482)					**Non-Transience Data Set** (Schools = 479)				
			Std'zed	Metric	SE	T-ratio	P-value	Std'zed	Metric	SE	T-ratio	P-value
INTRCPT1,												
INTRCPT2,												
	INTRCPT3	G000	1.35	1.35	0.01	135.25	0.00 *kj*	1.39	1.39	0.01	133.60	0.00 *kj*
	Y3NSCO_2	G001	0.07	0.12	0.03	3.89	0.00	×××	×××	×××	×××	×××
	ABSENT_2	G002	-0.13	-2.48	0.54	-4.61	0.00	-0.12	-2.38	0.48	-4.92	0.00
	PSCARD_2	G003	-0.08	-0.44	0.08	-5.49	0.00	-0.11	-0.60	0.07	-8.37	0.00
OCC		B01	-0.02	-0.02	0.01	-3.51	0.00 *j*	-0.02	-0.02	0.01	-3.12	0.00 *j*
AGE_1		B02	-0.03	-0.20	0.08	-2.50	0.01	×××	×××	×××	×××	×××
HOME_1		B03	0.05	0.16	0.07	2.29	0.02	×××	×××	×××	×××	×××
SEX												
INTRCPT2,												
	INTRCPT3	G100	-0.04	-0.09	0.01	-10.01	0.00 *j*	-0.05	-0.10	0.01	-10.62	0.00 *j*
	Y3NSCO_2	G101	-0.01	-0.04	0.02	-2.02	0.04	×××	×××	×××	×××	×××
AGE		B20	-0.07	-0.19	0.01	-15.57	0.00	-0.07	-0.19	0.01	-14.37	0.00
ATSI		B30	0.04	0.19	0.02	7.90	0.00	0.03	0.20	0.03	7.15	0.00
INOZ		B40	-0.01	-0.04	0.01	-2.83	0.01	-0.01	-0.04	0.01	-2.91	0.00
TRANS												
INTRCPT2,												
	INTRCPT3	G500	-0.04	-0.11	0.01	-8.33	0.00 *j*					
	AGE_2	G501	-0.02	-0.48	0.24	-2.06	0.04					
	Y3NSCO_2	G502	0.03	0.13	0.03	4.35	0.00					
Y3NSCORE,												
INTRCPT2		B60	0.69	0.55	0.01	115.46	0.00 *k*	0.70	0.57	0.01	116.82	0.00 *j*
INOZ_1		B61	0.01	0.11	0.05	2.36	0.02	×××	×××	×××	×××	×××

(Continued)

Table 9.4 Final estimation of fixed effects from Type B effects models (*Continued*)

2) Literacy

			Transience Data Set (Schools = 482)					Non-Transience Data Set (Schools = 479)				
			Std'zed	Metric	SE	T-ratio	P-value	Std'zed	Metric	SE	T-ratio	P-value
INTRCPT1,												
INTRCPT2,												
	INTRCPT3	G000	1.46	1.46	0.01	170.87	0.00 *kj*	1.51	1.51	0.01	152.41	0.00 *kj*
	Y3LSCO_2	G001	0.05	0.08	0.03	3.06	0.00	×××	×××	×××	×××	×××
	ABSENT_2	G002	-0.11	-2.10	0.32	-6.55	0.00	-0.10	-2.03	0.69	-2.92	0.00
	PSCARD_2	G003	-0.05	-0.30	0.07	-4.48	0.00	-0.07	-0.39	0.07	-6.03	0.00
	INOZ_2	G004	-0.02	-0.28	0.13	-2.23	0.03	×××	×××	×××	×××	×××
OCC		B01	-0.11	-0.10	0.01	-16.21	0.00 *j*	-0.11	-0.10	0.01	-15.69	0.00 *j*
SEX		B10	0.02	0.04	0.01	5.27	0.00 *j*	0.02	0.03	0.01	4.03	0.00 *j*
AGE		B20	-0.06	-0.18	0.01	-17.65	0.00	-0.06	-0.17	0.01	-15.25	0.00
ATSI		B30	0.03	0.15	0.02	7.53	0.00	0.03	0.14	0.03	5.40	0.00
HOME		B40	0.02	0.03	0.01	3.92	0.00	0.01	0.02	0.01	2.61	0.01
INOZ		B50	-0.02	-0.05	0.01	-4.93	0.00	-0.02	-0.05	0.01	-4.43	0.00
TRANS		B60	-0.02	-0.06	0.01	-5.52	0.00					
Y3LSCORE		B70	0.77	0.61	0.00	145.19	0.00 *kj*	0.78	0.62	0.00	139.19	0.00 *kj*

Notes: Shade - The variable TRANS is not available for examination in this model.
Std'zed - Regression coefficient obtained using standardized variables.
Metric - Regression coefficient obtained using unstandardized variables.
k - Residual parameter of this coefficient is left to vary at the occasion-level.
j - Residual parameter of this coefficient is left to vary at the school-level.
- Standard errors (SE), t-ratios and p-values presented are those obtained using unstandardized variables.

At the macro-level, the results for Type B effects (Table 9.4) show that there are statistically significant effects of the Average Proportion of School Cardholders (PSCARD_2) and of Average Absenteeism Rate (ABSENT_2) on student achievement in both numeracy and literacy regardless of the data set used. Clearly, high rates of absenteeism and low social economic status are negatively related to each of the two outcome measures. Surprisingly, however, the variable MOBILI_2 (Average Mobility Rate) has no significant effects on any of the two outcome measures. These results seem to be contrary to those of the analyses presented in the preceding chapters that indicate that the variable MOBILITY (Mobility Rate of a school within a testing occasion) has a significant influence on achievement in numeracy and literacy. However, it should be borne in mind that the models specified in this chapter (except at Level-1) are different from the models specified in the preceding chapters. Thus, it appears that the findings of this study indicate that, with appropriate modelling of the time variable (OCC) and with all significant student-level and school-level factors considered, the average mobility rate of a school has no significant ($p<0.05$) influence on achievement in numeracy and literacy.

In addition, at the macro-level, considering the transience data set, the results for Type B effects show that Average Prior Achievement over the study period (Y3NSCO_2 or Y3LSCO_2) of the school is positively related to each of the two outcome measures. That is, students in schools with high Average Prior Achievement scores are likely to perform better on the tests than their counterparts in schools which have low Average Prior Achievement scores. The results here also indicate that the variable INOZ_2 (overall Average Living in Australia) has a significant effect on achievement in literacy but this is not a consistent finding across the two data sets.

For numeracy, when considering the transience data set, the results in Table 9.4 indicate the existence of cross-level interaction effects between (a) Y3NSCORE and INOZ_1, (b) SEX and Y3NSCO_2, (c) TRANS and AGE_2, and (d) TRANS and Y3NSCO_2. Generally, these cross-level interaction effects are similar to the ones encountered in the analyses reported in the preceding chapters.

Finally, it should be noted from Table 9.4 that there are no measures of average school-context that have significant effects on the occasion slope. Therefore, the variability in the change effects can not be explained using the available measures of school-context.

Stable and change variance components

Tables 9.5 and 9.6 present the results of the estimation of the variance components from the Type A and Type B effects models respectively. Each table presents the values of the variance components of the stable school effects and the variance components of the change school effects for the two outcome measures and using the two data sets. The variance components associated with all the variables with random effects at the second and the third levels of the simplest longitudinal models, the Type A effect models and the Type B effect models can be found in Appendices 14.4 and 14.5.

Clearly, based on the chi-square statistics and the p-values associated with each stable and each change component of school effects presented in Tables 9.5 and 9.6, all the components presented in the two tables are statistically significant at the 0.05 probability level. Thus, the results in the Tables 9.5 and 9.6 indicate that the primary schools in South Australia are different in terms of the stable (average effectiveness) and change (improvement or deterioration) Type A as well as Type B school effects for numeracy and literacy.

Table 9.5 Final estimation of variance components from Type A effects models

		Transience Data Set (Schools = 482)					Non-Transience Data Set (Schools = 479)				
		Var. Comp.	df	Chi-Square	P-Value	Stability Ratio	Var. Comp.	df	Chi-Square	P-Value	Stability Ratio
Numeracy											
	Stable ($\mathbf{u}_{00j}$)	0.045	462	1369.68	0.00	**12.20**	0.041	446	1184.68	0.00	**10.42**
	Change ($\mathbf{u}_{01j}$)	0.004	462	594.43	0.00		0.004	446	563.11	0.00	
Literacy											
	Stable ($\mathbf{u}_{00j}$)	0.027	358	755.30	0.00	**8.10**	0.022	335	593.75	0.00	**7.17**
	Change ($\mathbf{u}_{01j}$)	0.003	358	440.14	0.00		0.003	335	391.36	0.02	

Table 9.6 Final estimation of variance components from Type B effects models

		Transience Data Set (Schools = 482)					Non-Transience Data Set (Schools = 479)				
		Var. Comp.	df	Chi-Square	P-Value	Stability Ratio	Var. Comp.	df	Chi-Square	P-Value	Stability Ratio
Numeracy											
	Stable ($\mathbf{u}_{00j}$)	0.018	459	842.30	0.00	**4.94**	0.020	460	867.62	0.00	**5.17**
	Change ($\mathbf{u}_{01j}$)	0.004	462	593.00	0.00		0.004	462	595.07	0.00	
Literacy											
	Stable ($\mathbf{u}_{00j}$)	0.014	462	816.97	0.00	**4.79**	0.013	460	740.11	0.00	**4.73**
	Change ($\mathbf{u}_{01j}$)	0.003	466	582.20	0.00		0.003	462	570.56	0.00	

The stability ratios (given in **bold** in Tables 9.5 and 9.6) are obtained by dividing the stable variance component of the school effect ($\mathbf{u}_{00j}$) with the change variance component of the school effect ($\mathbf{u}_{01j}$). Willms and Raudenbush (1989) have argued that these stability ratios provide information regarding the magnitudes of differences between schools in their stable components relative to the magnitude of the change components; with low ratios being associated with less stable effects.

Using the stability ratio criterion, clearly, the Type B effects are less stable than Type A effects. However, when the same type of school effects are compared across the two outcome measures as well as across the two data sets, it appears that the stability of the school effects do not differ considerably. Nevertheless, it should be noted that the stability ratios for numeracy are in general slightly higher than the stability ratios for literacy, which seems to suggest that the school effects for numeracy are marginally more stable than those for literacy.

Variance partitioning and variance explained

In order to give a proper account of the amounts of variance involved in the analyses presented above, the simplest model for the longitudinal estimation of variation among school effects should include both the stable component and the component that varies across occasion. Consequently, the simplest model has the time trend variable OCC as the only predictor and no other predictor variables are specified at any level of this model. Hence, employing the notation introduced above for Type A

and Type B models, the simplest model for the longitudinal estimation of variation among school effects is as follows.

Level-1 model

$$\mathbf{Y}_{itj} = \pi_{0tj} + \mathbf{e}_{itj}$$

Level-2 model

$$\pi_{0tj} = \beta_{00j} + \beta_{01j}\mathbf{OCC}_{tj} + \mathbf{r}_{0tj}$$

Level-3 model

$$\beta_{00j} = \gamma_{000} + \mathbf{u}_{00j}$$

$$\beta_{01j} = \gamma_{010} + \mathbf{u}_{01j}$$

Equation 9.18

The time trend variable OCC in Equation 9.18 is group-mean centred in these analyses (Kreft, 1995; Kreft et al., 1995). In addition, all the components in Equation 9.18 carry the same meaning as described above for models for Type A and Type B effects.

Tables 9.7 and 9.8 give estimates of variances involved in the Type A and Type B effects models, respectively. Rows 'a' and 'b' show the variance components obtained from the simplest longitudinal models and the variance components obtained from the final school effects models, respectively. The entries in rows 'c' to 'f' of Tables 9.7 and 9.8 are calculated from the results in rows 'a' and 'b' of the tables following the procedure described in Chapter 4.

Thus, the results in Tables 9.7 and 9.8 show that the percentages of variance explained at Levels 2 and 3 are very large (over 80 per cent) compared to the percentages of variances explained at Level-1 (between 44 and 56 per cent), regardless of the type of school effects considered. Consequently, the percentages of total variances left unexplained (especially at the school-level, which is Level-3) are small (less than three per cent). These results are consistent across the two subjects (numeracy and literacy) and across the two data sets used.

However, from the results in Tables 9.7 and 9.8, it should be noted that less variance is left unexplained at the school-level in Type B effects models compared to what is left unexplained at that level in Type A effects models. For example, for numeracy when considering the transience data set, the percentage of variance left unexplained in Type A effects models is 2.9 per cent, and in Type B effects model, it is 1.1 per cent. It should further be noted that in Type A effects models, the percentages of variance left unexplained at either Level-1 or Level-2 are mostly equal to the percentages of variance left unexplained at the same level in Type B effects models. For example, for numeracy when considering the non-transience data set, the percentages of variance left unexplained in Levels 1 and 2 of Type A effects model are 37.6 and 2.9 per cent respectively, which are the same as in Type B effect models.

In interpreting the results in Tables 9.7 and 9.8, it must be borne in mind that Raudenbush and Willms (1995) cautioned against assuming that the amount of variation between schools puts an upper limit on the variation of school effects (either Type A or Type B). Raudenbush and Willms argued that the variation attributed to either type of school effect could be larger than the overall variation between schools for a number of reasons, one of the reasons being that either type of school effect can influence within school variation by interacting with student background.

Finally, it is worth noting that, on the whole, the percentages of total variance explained based on the three-level longitudinal models reported in this chapter are

noticeably larger than the corresponding percentages that are explained based on the two-level and three-level models reported in Chapters 7 and 8 respectively. For example, for numeracy and based on the three-level analyses reported in Chapter 8 (Table 8.6), the percentages of total variance explained are 49.5 and 49.8 using the transience and the non-transience data sets respectively. The results in Tables 9.7 and 9.8 show that corresponding percentages based on the analyses reported in this chapter are noticeably larger for Type A (58.5 and 56.7) as well as for Type B (60.3 and 58.1) models for the transience and non-transience data sets respectively. Thus, if the total amounts of variance explained were to be used as a measure of how good a model fits the data, then the models under the longitudinal design are better models compared to the two-level and three-level models discussed in the preceding chapters. Arguably, there is an improved fit of the model to the data under the longitudinal structure because the time variable (OCC) is modelled better under the longitudinal structure than under the two- or three-level structures employed in the earlier chapters.

Correlations

The next three sub-sections focus on the correlations[27] among the schools effects computed above. The first two sub-sections focus on the correlations between the indices of individual school effects computed above, while the third sub-section addresses the question of consistency of school effects across occasions.

The correlation coefficients provided in this study are for empirical Bayes (EB) estimates of individual school effects, not for ordinary least square (OLS) estimates. Therefore, these coefficients should be interpreted with some caution because Willms and Raudenbush (1989) reported that EB estimates do exaggerate the stability of school effects. However, Willms and Raudenbush (1989; p. 232) showed that "although the EB estimates do exaggerate the stability of school effects, they supply a much more credible picture of the distribution of school effects than do the OLS estimates".

Cohen (1992; p.157) suggests that correlation coefficients below |0.10| are trivial, coefficients between |0.10| and |0.29| are “small”, coefficients between |0.30| and |0.49| are “medium” or “moderate”, and coefficients above |0.50| are “large” or “strong”. In addition, in this study, correlation coefficients between |0.60| and |0.79| are termed ‘very large’ or ‘very strong’, and those above |0.80| are termed ‘extremely large’ or ‘extremely strong’.

For purposes of ease in presentation, codes are used to name the school effects in this chapter. School effects obtained using the transience data set have the prefix 'T' at the beginning of their codes, and those obtained using the non-transience data set have a 'V'. The prefixes 'EB00' and 'EB01' represent the empirical Bayes estimates for the stable and the change (school-by-occasion) components of school effects, respectively. Codes for school effects for numeracy have the prefix 'N' while literacy have 'L'. Likewise, codes for Type A effects have the suffix 'A' at the end while for Type B effects have the suffix 'B'. For example, in Table 9.9, the code 'TEB00NA' represents the stable Type A school effects for numeracy obtained using all the students who could be matched, and the code 'VEB00LB' represents the stable Type B school effects for literacy obtained using those students who remained in the same school.

[27] the correlation coefficients reported in this chapter are computed using *SPSS 10.0.5 for Windows*.

Table 9.7 Longitudinal estimation of variation among school Type A effects

	Transience Data Set (Schools = 482)				**Non-Transience Data Set** (Schools = 479)			
	Level-1	Level-2	Level-3	Total	Level-1	Level-2	Level-3	Total
	(N=37,832)	(N=1,853)	(N=482)		(N=32,741)	(N=1,823)	(N=479)	
Numeracy								
a) Var. Comp. Simplest Longitudinal Model	1.00	0.29	0.25	1.54	0.99	0.23	0.22	1.43
b) Var. Comp. Type A Effects Model	0.56	0.04	0.05		0.54	0.04	0.04	
c) Var. Available	65.0%	19.1%	16.0%		68.8%	16.0%	15.2%	
d) Var. Explained	44.2%	87.6%	81.8%		45.4%	82.1%	81.1%	
e) Total Var. Explained	28.7%	16.7%	13.1%	58.5%	31.2%	13.1%	12.3%	56.7%
f) Var. Left Unexplained	36.3%	2.4%	**2.9%**	41.5%	37.6%	2.9%	**2.9%**	43.3%
Literacy								
a) Var. Comp. Simplest Longitudinal Model	1.00	0.23	0.22	1.44	0.98	0.21	0.19	1.37
b) Var. Comp. Type A Effects Model	0.45	0.03	0.03		0.43	0.04	0.02	
c) Var. Available	69.3%	15.8%	15.0%		71.3%	15.2%	13.5%	
d) Var. Explained	55.1%	85.1%	87.6%		55.9%	82.3%	88.0%	
e) Total Var. Explained	38.2%	13.4%	13.1%	64.7%	39.8%	12.5%	11.9%	64.2%
f) Var. Left Unexplained	31.1%	2.4%	**1.9%**	35.3%	31.4%	2.7%	**1.6%**	35.8%

Table 9.8 Longitudinal estimation of variation among school Type B effects

	Transience Data Set (Schools = 482)				**Non-Transience Data Set** (Schools = 479)			
	Level-1	Level-2	Level-3	Total	Level-1	Level-2	Level-3	Total
	(N=37,832)	(N=1,853)	(N=482)		(N=32,741)	(N=1,823)	(N=479)	
Numeracy								
a) Var. Comp. Simplest Longitudinal Model	1.00	0.29	0.25	1.54	0.99	0.23	0.22	1.43
b) Var. Comp. Type A Effects Model	0.56	0.04	0.02		0.54	0.04	0.02	
c) Var. Available	65.0%	19.1%	16.0%		68.8%	16.0%	15.2%	
d) Var. Explained	44.2%	87.8%	92.8%		45.3%	81.7%	90.7%	
e) Total Var. Explained	28.7%	16.7%	14.8%	60.3%	31.2%	13.1%	13.8%	58.1%
f) Var. Left Unexplained	36.2%	2.3%	**1.1%**	39.7%	37.6%	2.9%	**1.4%**	41.9%
Literacy								
a) Var. Comp. Simplest Longitudinal Model	1.00	0.23	0.22	1.44	0.98	0.21	0.19	1.37
b) Var. Comp. Type A Effects Model	0.45	0.03	0.01		0.43	0.04	0.01	
c) Var. Available	69.3%	15.8%	15.0%		71.3%	15.2%	13.5%	
d) Var. Explained	55.0%	85.0%	93.6%		55.7%	82.2%	92.7%	
e) Total Var. Explained	38.1%	13.4%	14.0%	65.5%	39.7%	12.5%	12.5%	64.8%
f) Var. Left Unexplained	31.2%	2.4%	**1.0%**	34.5%	31.5%	2.7%	**1.0%**	35.2%

Correlations between Types A and B school effects

The top panel of Table 9.9 shows the correlations between the stable Types A and B effects obtained using the transience data set and those obtained using the non-transience data set, while the bottom panel of the table shows the corresponding information for change Types A and B effects.

Table 9.9 Correlations between school effects across data sets and across outcome measures

		Numeracy		Literacy	
		Tran$^{\propto}$	Non-Tran$^{\beta}$	Tran$^{\propto}$	Non-Tran$^{\beta}$
Stable**					
	Type A				
		TEB00NA	VEB00NA	TEB00LA	VEB00LA
	TEB00NA	1.00			
	VEB00NA	**0.97**	1.00		
	TEB00LA	0.71	0.67	1.00	
	VEB00LA	0.68	0.69	**0.96**	1.00
	Type B				
		TEB00NB	VEB00NB	TEB00LB	VEB00LB
	TEB00NB	1.00			
	VEB00NB	**0.98**	1.00		
	TEB00LB	0.50	0.50	1.00	
	VEB00LB	0.54	0.56	**0.94**	1.00
Change**					
	Type A				
		TEB01NA	VEB01NA	TEB01LA	VEB01LA
	TEB01NA	1.00			
	VEB01NA	**0.92**	1.00		
	TEB01LA	0.18	0.13	1.00	
	VEB01LA	0.16	0.15	**0.94**	1.00
	Type B				
		TEB01NB	VEB01NB	TEB01LB	VEB01LB
	TEB01NB	1.00			
	VEB01NB	**0.95**	1.00		
	TEB01LB	0.28	0.24	1.00	
	VEB01LB	0.24	0.24	**0.94**	1.00

Notes: $\propto$ - Using all the students matched (Schools = 482).
β - Using only those students matched in the same school (Schools = 479).
** - All the correlations are significant at the 0.01 level.

The numbers given in bold in Table 9.9 are the coefficients of the correlations between the stable (or between the change) Type A (or Type B) effects obtained using the transience data set and the ones obtained using the non-transience data set within the same subject area. Thus, within the same type of effect (stable or change), the results in Table 9.9 show extremely strong correlations (near unity) within the same subject between the Type A (or Type B) school effects obtained using the transience data set and the school effects obtained using the non-transience data set. Obviously,

based on either Type A or Type B effects, the ranking of schools obtained using all students who could be matched and the ranking obtained using the students who remained in the same school do not differ markedly.

For the stable school effects, the results in the top panel of Table 9.9 show very strong correlations (0.67 to 0.71) for Type A effects and large correlations (0.50 to 0.56) for Type B effects, across the two subjects regardless of the data set used. The very strong correlations between the stable Type A effects across the two subjects indicate that after controlling for student intake, a vast majority of schools that perform well in numeracy also perform well in literacy, and that a vast majority of schools that perform poorly in numeracy also perform poorly in literacy.

In other words, for stable Type A effects, there are many schools that show consistent performance across the two outcome measures, that is, stable Type A effects are to a great extent consistent across the two subjects included in the BSTP. However, the large correlations across the two subjects for Type B effects indicate that, after adjusting for student intake and school context, there are fewer schools that show consistent performance across the two subjects compared to the number of schools showing consistent performance when Type A effects are considered. That is, Type B effects are relatively less consistent across the two outcome measures compared to Type A effects, as might be expected (Willms and Raudenbush, 1989).

For the change school effects, the results in the bottom panel of Table 9.9 show small correlations for Type A effects (0.13 to 0.18) and also for Type B effects (0.24 to 0.28), across the two subjects regardless of the data set used. Thus, based on either Type A or Type B effects, only a small number of schools that record more than expected change in performance over time in numeracy also record more than expected change in performance over time in literacy and vice versa.

Table 9.10 displays the correlations between the Type A effects and Type B effects obtained using all the students who could be matched as well as those obtained using those students who remained in the same school. The figures given in bold in Table 9.10 are the correlations between the Type A and Type B school effects within the same subject for one data set.

Thus, the results in Table 9.10 show strong to very strong to extremely strong correlations (0.78 to 0.90) within the same subject between the stable Type A and Type B school effects within one data set as well as across the two data sets used. Likewise, these results show extremely strong correlations (0.85 to 0.98) between the change Type A and Type B effects. Clearly, most schools show consistent performance across Types A and B school effects, stable or change.

Correlations between stable and change school effects

Research on change has found that it is not possible to obtain a consistent estimate of the correlation between individual change and initial status in a simple pretest-posttest design. In particular, researchers have found that the measurement errors in the pretest and the observed change score are commonly negatively correlated and this leads to the spurious negative correlation typically found between the initial status and the rate of change (Bereiter, 1963; Blomqvist, 1977; Willett, 1988). However, Rogosa (1995; p. 17) argues that the correlation between change and initial status can be negative, zero or positive because the correlation "depends crucially on the choice of t_1, the time at which the initial status is measured". Nevertheless, Willett (1988), Bryk and Raudenbush (1992) and Muller, Stage and Kinzie (2001) contend that a consistent estimation of the correlation can be obtained with multiwave data (that is, a longitudinal design that incorporates data on successive cohorts of individuals).

Table 9.10 Correlation between Type A and Type B school effects

		Type A effects			
		Stable		Change	
		Tran∝	Non-Tranβ	Tran∝	Non-Tranβ
Type B effects	**Numeracy****				
		TEB00NA	VEB00NA	TEB01NA	VEB01NA
	TEB00NB	**0.78**	0.82		
	VEB00NB	0.79	**0.84**		
	TEB01NB			**0.98**	0.91
	VEB01NB			0.93	**0.96**
	Literacy**				
		TEB00LA	VEB00LA	TEB01LA	VEB01LA
	TEB00LB	**0.85**	0.83		
	VEB00LB	0.84	**0.90**		
	TEB01LB			**0.89**	0.85
	VEB01LB			0.85	**0.90**

Notes: ∝ - Using all the students matched (Schools = 482).
β - Using only those students matched in the same school (Schools = 479).
** - All the correlations are significant at the 0.01 level.

It should be remembered that a longitudinal design is employed in the estimation of school effects in this chapter. Under the longitudinal design "each school serves as its own control" (Willms and Raudenbush, 1989; p. 214). This is because data on successive cohorts of students are used to establish the progress of a school: not relative to the performance of the other schools in the state but relative to its own performance. In particular, the estimated effect of a school includes a stable component (that is, its average effect over the period of the study) and a change component (that is, an effect specific to each point time). However, the value of the 'school-by-occasion' interaction ($\mathbf{u}_{01j}\mathbf{OCC}_{ij}$) or 'change' effect is considered the important element of the change component of school effect because it shows the change (improvement or deterioration) that occurred in a school's Type A or Type B effect over the study period.

Consequently, this sub-section focuses on the relation between the average effect of a school and its 'school-by-occasion' effect. The question here is 'Do schools that show more than expected average performance also show more than expected increase in performance over time?'

Table 9.11 shows the correlation between the school effects EB00 (stable) and the school-by-occasion (or simply 'change') effects EB01 using the two data sets for each of the outcome measures. The figures given in bold in Table 9.11 are the correlations between stable school effects and change school effects within the same subject and for one data set.

For numeracy, the results in Table 9.11 show small to medium but positive correlations (0.27 to 0.43) between the stable effects and the change effects for Type A as well as for Type B effects. The positive correlations indicate that based on Type A or Type B measures of school effectiveness, a considerable number of schools that recorded more than expected average performance in numeracy also recorded more

than expected change in performance in numeracy over time. Alternatively, a considerable number of schools that recorded less than expected average performance in numeracy also recorded less than expected change in performance in numeracy over time. However, because some of these correlations are small, it indicates that schools are not highly consistent in terms of the relationship between their average performance and their change in performance over time. That is, schools that show more than expected average performance in numeracy do not inevitably exhibit more than expected change in performance in numeracy over time and vice versa. Nevertheless, most of these correlations are within the so-called "medium" (Cohen, 1992; p.157) correlations range, 0.30 to 0.49, (especially for Type B effects) which indicate existence of considerably consistent relationship between the stable school effects and the change school effects with respect to numeracy.

Table 9.11 Correlation between stable and change school effects

		Change effects			
		Type A		**Type B**	
		Tran$^{\propto}$	Non-Tran$^{\beta}$	Tran$^{\propto}$	Non-Tran$^{\beta}$
Stable effects	**Numeracy****				
		TEB01NA	VEB01NA	TEB01NB	VEB01NB
	TEB00NA	**0.27**	0.28	0.30	0.30
	VEB00NA	0.31	**0.29**	0.34	0.33
	TEB00NB	0.34	0.30	**0.43**	0.42
	VEB00NB	0.34	0.31	0.43	**0.43**
	Literacy**				
		TEB01LA	VEB01LA	TEB01LB	VEB01LB
	TEB00LA	**-0.84**	-0.82	-0.67	-0.69
	VEB00LA	-0.76	**-0.82**	-0.61	-0.71
	TEB00LB	-0.72	-0.72	**-0.75**	-0.76
	VEB00LB	-0.67	-0.74	-0.68	**-0.77**

Notes: $\propto$ - Using all the students matched (Schools = 482).
β - Using only those students matched in the same school (Schools = 479).
** - All the correlations are significant at the 0.01 level.

For literacy, the results in Table 9.11 show very strong to extremely strong but negative correlations (-0.61 to -0.84) between the stable school effects and the change school effects for Type A as well as for Type B effects. The negative correlations indicate that based on Type A or Type B measures of school effectiveness, most schools that recorded more than expected average performance in literacy recorded less than expected increase in performance in literacy over time, and vice versa. Because these correlations are strong, it indicates that this inverse relationship between the stable school effects and the change school effects is decidedly consistent. That is, when things are equal, primary schools in South Australia that have more than expected average performance in literacy almost certainly exhibit less than expected increase in performance in literacy over time. Alternatively, the schools that have less than expected average performance in literacy almost certainly exhibit more than expected increase in performance in literacy over time.

One issue comes out clearly from the results in Table 9.11: the relationships between the stable school effects and the change school effects differ for the two outcome

measures. Whereas the relationship is positive for numeracy, it is decidedly negative for literacy. It should be considered that the longitudinal structure employed here overcomes the problem of spurious results due to negative correlation between the measurement errors in the initial status (in this case, average performance) and true change (in this case, increase in performance) (Bryk and Raudenbush, 1992). With longitudinal structure, negative correlations between initial status and change in academic achievement are not strange findings because some researchers have reported such correlations though not as strong as found here for literacy (e.g. see Raudenbush and Bryk 1988, p.462 and 1992, p.138; Willms and Raudenbush 1989, p. 223; and Embretson, 1995; p. 196). Consequently, for literacy, it can be inferred that there is a clear negative relationship between the stable school effects and the change school effects.

Nevertheless, the results in Table 9.11 do raise an interesting question. Why is the relationship between the stable school effects and the change school effects for numeracy found to be completely different from that of literacy? It should be borne in mind that all analyses carried out thus far have produced more or less identical findings regarding the two outcome measures. It should also be borne in mind that identical procedures were followed to obtain the scores for the two outcome measures. Consequently, any errors in the outcome variables introduced in equating of the tests should be expected to affect the estimation of the school effects for the two outcome measures to roughly the same extent, and probably, in the same direction. Furthermore, any measurement errors in the predictor variables and any errors due to differential participation rates should be expected to affect the results in a similar direction because the same data sets are used to estimate school effects for the two outcome measures.

It is obvious that, based on the analyses carried out thus far, there is no clear answer to the above issue. Consequently, further analyses are undertaken to examine the data more carefully in attempts to provide an answer to the question. The purposes of these analyses include looking for:

(a) differences in score distribution between the two outcome measures;

(b) the direction of the correlation coefficient between the stable school effects and the change school effects in the null model;

(c) the direction of the correlation coefficient between the stable school effects and the change school effects when MLwiN (Browne et al., 2001) software is employed instead of the HLM5/3L (Raudenbush et al., 2000) software; and

(d) the direction of the correlation coefficient between the stable school effects and the change school effects if the literacy test is broken down into its sub-tests, that is, language and reading.

The analyses undertaken are discussed in Appendix 14.6. By and large, all the above attempts [(a) to (d)], made to provide an explanation to the results in Table 9.11, were fruitless and, therefore, it is unclear why the relationship between the stable school effects and the change school effects for numeracy is different from that of literacy.

First, the frequency distributions of scores for the two outcome measures are found to be consistently similar, which suggests that the chances of a ceiling effect in one of the outcome variables is unlikely. However, there are some indications that the values of skewness of the distributions for literacy scores have increased over time. Nevertheless, the observed change in skewness values per year is very small (0.09) and, therefore, the evidence is considered insufficient for making sound conclusions regarding the existence of a ceiling effect in the literacy tests.

Second, it is found that the control of the student background characteristics does not affect the direction of the correlation between the stable school effects and the change school effects for the two outcome variables. Therefore, it is concluded that the results in Table 9.11 above do not arise wholly from differences in the nature of the contribution made by the student background factors to each of the two outcome measures.

Third, it is found that the direction of correlations between the stable school effects and the change school effects for reading and language is the same as the direction obtained for literacy, which indicate that the dimensionality of the literacy test does not explain the results in Table 9.11. That is, it is appropriate to combine the reading and the language sub-scales to form a single literacy scale.

Finally, it is found that the direction of correlation between the two components of school effects for numeracy and literacy is positive when the Grade 3 scores (rather than Grade 5 scores) are used as the outcome variables in the simplest longitudinal model. For literacy, this indicates that there are some major shifts in the relationship between the stable school effects and the change school effects somewhere in-between the two grades, but do not explain the results in Table 9.11.

The next sub-section examines the consistency of school effects across the occasions. Only Type A effects are considered in the next sub-section because the structure of the model for estimating Type B effects for each cohort of students differs from the longitudinal model used to estimate the stable Type B effects.

Consistency of Type A effects across the occasions

This third sub-section examines the consistency of Type A effects by comparing the correlations between the school-level residuals obtained using the four cohorts of students; namely 1995/1997, 1996/1998, 1997/1999 and 1998/2000 cohorts. The sub-section also examines the correlations between the stable components of Type A effects estimated from the three-level longitudinal model (Equation 9.8) and the Type A effects estimated using each of the four cohorts of students. Two-level models that treat each school as a separate entity on each of the four testing occasions (as described in Chapter 7) are employed to estimate Type A effects for the four cohorts of students. In the two-level model unique identity has to be used for each Level-2 unit (school-level), and therefore, different identities are used to represent each school on the various occasions, which makes it possible to estimate school effects for each cohort simultaneously.

For presentation purposes, the same codes used to name school effects in the previous sub-sections are used in this sub-section. However, in order to differentiate between the school effects for each of the four occasions, numbers are incorporated in the codes. For example, in Table 9.12, the code 'TEB95NA' and 'TEB96NA' represent the Type A school effects for numeracy obtained using all the students who could be matched for the 1995/1997 and 1996/1997 cohorts, respectively.

Estimation of Type A effects for each occasion

It is mentioned above that to estimate Type A effects, two-level models that control for the student intake (as in the three-level models) are estimated for each outcome measure. The estimations are carried out first using the transience data set and then the non-transience data set. Hence, the two-level models employed to estimate Type A effects in this sub-section are identical to the three-level longitudinal model at Levels 1 and 3.

For example, the two-level model for estimation of Type A effects for numeracy using all the students who could be matched is as follows:

Level-1 model

$$Y = B0 + B1*(SEX) + B2*(TRANS) + B3*(AGE) + B4*(ATSI) + B5*(INOZ) + B6*(Y3NSCORE) + R$$

Level-2 model

$$B0 = G00 + U0$$

$$B1 = G10 + U1$$

$$B2 = G20 + U2$$

$$B3 = G30$$

$$B4 = G40$$

$$B5 = G50$$

$$B6 = G60$$

Equation 9.19

It should be noted that the student background characteristics included in Equation 9.19 are the same ones included in the three-level longitudinal model (Equation 9.16). In addition, the effects of student-level variables that are modelled as fixed at Level-3 in the three-level model (Equation 9.16) are also modelled as fixed at Level-2 in the two-level model presented above. That is, except for the effects of SEX (Sex of the Student) and TRANS (Transience), the effects of all the other student-level variables are modelled as fixed at Level-2 of Equation 9.19.

The above approach is employed to estimate the Type A effects for numeracy and literacy using the two data sets.

Correlations between overall Type A effects and Type A effects for each occasion

Table 9.12 shows the correlations between the stable (overall) components of Type A effects estimated from the three-level longitudinal models (EB00) and the Type A effects estimated from the two-level models using each of the four cohorts of students (EB95 to EB97) for the two outcome measures. The correlation coefficients presented in Table 9.12 should be interpreted with some caution because the overall school effects also contain the effects from each of the four cohorts, and therefore, the correlations are exaggerated to some extent. Nevertheless, the results in Table 9.12 do supply a general picture of the consistency of school effects.

For numeracy, the results in Table 9.12 show very strong correlations (0.60 to 0.75) between the overall Type A effects and the Type A effects for each cohort regardless of the data set used, which indicate that the school effects for each cohort consistently agree with the overall school effects. However, the correlations for literacy are medium to very strong (0.47 to 0.77), which indicate that the school effects for some cohorts do not agree with the overall school effects.

Correlations between Type A effects from each testing occasion

Table 9.13 shows the correlations between Type A effects estimated from the two-level models using the four cohorts of students for numeracy as well as literacy. The figures given in bold in Table 9.13 are the correlations between school effects

obtained using the transience data set and the school effects obtained using the non-transience data set within the same subject and on the same testing occasion. Obviously, for both outcome measures and for each of the four cohorts, the rank of a school obtained using all students who could be matched and the rank obtained using the students who remained in the same school do not differ markedly ($r \geq 0.95$).

Table 9.12 Correlations between overall Type A effects and Type A effects for each occasion

		Stable Type -A effects	
		Tran$^{\propto}$	Non-Tran$^{\beta}$
Type A effects for each occasion**	**Numeracy**		
		TEB00NA	VEB00NA
	TEB95NA	0.61	
	TEB96NA	0.69	
	TEB97NA	0.75	
	TEB98NA	0.69	
	VEB95NA		0.60
	VEB96NA		0.68
	VEB97NA		0.74
	VEB98NA		0.68
	Literacy		
		TEB00LA	VEB00LA
	TEB95LA	0.77	
	TEB96LA	0.68	
	TEB97LA	0.60	
	TEB98LA	0.47	
	VEB95LA		0.76
	VEB96LA		0.65
	VEB97LA		0.58
	VEB98LA		0.47

Notes: $\propto$ - Using all the students matched (Schools = 482).
β - Using only those students matched in the same school (Schools = 479).
** - All the correlations are significant at the 0.01 level.

For numeracy as well as literacy, the results in Table 9.13 show small to medium but positive correlations (0.16 to 0.40) between the Type A effects estimated from the two-level models using the four cohorts of students. The positive correlations indicate that a considerable number of schools that show more than expected performance on one testing occasion also show more than expected performance on the other testing occasions and vice versa. However, because some of these correlations are small, it indicates that schools are not highly consistent in terms of the relationship between their performance on one testing occasion and their performance on the other testing occasions. Thus, ranking of schools based on data on a single cohort of students could be very misleading.

The figures given in *italics* in Table 9.13 are the correlations between the school effects across the two outcome measures. The *italic* figures in the shaded cells of Table 9.13 are the correlations between school effects for numeracy and literacy within the same testing occasion. Clearly, only a small number of schools that show more than expected performance in numeracy on one testing occasion also show more than expected performance in literacy on the other testing occasions (r =0.11 to 0.34). However, within the same testing occasion, the results indicate that a considerable number of schools that show more than expected performance in numeracy also show more than expected performance in literacy regardless of the data set used (r = 0.54 to 0.64).

Table 9.13 Correlations between Type A effects estimated from two-level models for the four cohorts of students

	Numeracy**								**Literacy****							
	Transience Data Set$^{\propto}$				Non-Transience Data Set$^{\beta}$				Transience Data Set$^{\propto}$				Non-Transience Data Set$^{\beta}$			
	TEB95NA	TEB96NA	TEB97NA	TEB98NA	VEB95NA	VEB96NA	VEB97NA	VEB98NA	TEB95LA	TEB96LA	TEB97LA	TEB98LA	VEB95LA	VEB96LA	VEB97LA	VEB98LA
TEB95NA	1.00															
TEB96NA	0.37	1.00														
TEB97NA	0.26	0.35	1.00													
TEB98NA	0.20	0.26	0.40	1.00												
VEB95NA	**0.98**	0.35	0.24	0.18	1.00											
VEB96NA	0.37	**0.97**	0.33	0.26	0.35	1.00										
VEB97NA	0.23	0.34	**0.98**	0.39	0.22	0.33	1.00									
VEB98NA	0.18	0.25	0.39	**0.97**	0.16	0.24	0.38	1.00								
TEB95LA	*0.64*	*0.31*	*0.25*	*0.21*	*0.60*	*0.30*	*0.21*	*0.19*	1.00							
TEB96LA	*0.21*	*0.63*	*0.29*	*0.20*	*0.20*	*0.58*	*0.27*	*0.19*	0.28	1.00						
TEB97LA	*0.23*	*0.25*	*0.59*	*0.30*	*0.21*	*0.23*	*0.58*	*0.30*	0.25	0.27	1.00					
TEB98LA	*0.15*	*0.18*	*0.34*	*0.57*	*0.14*	*0.20*	*0.32*	*0.55*	0.19	0.17	0.31	1.00				
VEB95LA	*0.63*	*0.30*	*0.25*	*0.22*	*0.61*	*0.29*	*0.21*	*0.20*	**0.97**	0.27	0.25	0.18	1.00			
VEB96LA	*0.22*	*0.60*	*0.27*	*0.21*	*0.22*	*0.58*	*0.26*	*0.20*	0.27	**0.96**	0.24	0.18	0.26	1.00		
VEB97LA	*0.19*	*0.24*	*0.55*	*0.28*	*0.18*	*0.22*	*0.57*	*0.28*	0.20	0.25	**0.95**	0.29	0.20	0.23	1.00	
VEB98LA	*0.12*	*0.16*	*0.29*	*0.54*	*0.11*	*0.16*	*0.27*	*0.55*	0.17	0.16	0.28	**0.95**	0.16	0.16	0.26	1.00

Notes: $\propto$ - Using all the students matched (Schools = 482).
β - Using only those students matched in the same school (Schools = 479).
** - All the correlations are significant at the 0.01 level.

Conclusions and discussion

The first five sections of this chapter explore a longitudinal multilevel regression model approach for examining performance of the primary schools in South Australia over time using the scores from the BSTP. The longitudinal design is powerful because it separates the stable effects of schools from the changing components of their effects.

The results of the longitudinal model HLM analyses can be summarized as follows for the two outcome measures of interest in this study.

1. Stable Type A effects (λ = 0.46 to 0.64) are estimated more reliably than are the stable Type B (λ = 0.37 to 0.45) effects.
2. The stable components of either Type A or Type B effects (λ = 0.37 to 0.64) are estimated far more reliably than the change component of school effects (λ = 0.15 to 0.19).
3. The stable school effects for numeracy (λ = 0.44 to 0.64) are estimated slightly more reliably than for literacy (λ = 0.37 to 0.53).
4. There is significant ($p<0.05$) variability in effectiveness of the primary schools in South Australia based on Type A and Type B (both stable and change) indices of school effects.
5. Based on the stability ratio criterion (Willms and Raudenbush, 1989);
   i) Type B effects (stability ratio = 4.73 to 5.17) are less stable than Type A effects (stability ratio = 7.17 to 12.20) , and
   ii) Type A school effects for numeracy (stability ratio = 10.42 to 12.20) are marginally more stable than for literacy (stability ratio = 7.17 to 8.10), and the Type B for numeracy (stability ratio = 4.94 to 5.17) are marginally more stable than for literacy (stability ratio = 4.73 to 4.79).
6. Generally, the performance of the schools on the earlier occasions is estimated to be higher than their performance on the later occasions.
7. In terms of fixed effects, the results presented in this chapter strongly agree with the results of analyses presented in the earlier chapters regarding the student-level and school-level variables that have significant influences on achievement in numeracy and literacy.
8. The percentages of total variances left unexplained at the school-level are very small (between 1.0 and 2.9 per cent) compared to the proportions of variances left unexplained at the student-level (between 31.1 and 37.6 per cent) regardless of the type of school effects considered. Moreover, a little less variance is left unexplained at the school-level in Type B effects models compared to what is left unexplained at that level in Type A effects models.

The last section of this chapter focuses on the consistency of school effects across outcome measures and across occasions. The main findings in this section can be summarized as follows for the two outcome measures of interest in this study.

9. Based on the Type A or Type B school effects for numeracy and literacy, the ranking of schools obtained using all students who could be matched and the ranking obtained using the students who remained in the same school do not differ markedly ($r \geq 0.94$).

10. The stable Type A effects are to a great extent consistent (r = 0.67 to 0.71) across the two subjects included in the BSTP. However, the stable Type B effects are less consistent (r = 0.50 to 0.56) across the two outcome measures.
11. Within the same subject, a vast majority of the primary schools that perform well based on stable Type A effects also perform well based on Type B effects. That is, within the same subject Type A and Type B school effects are highly consistent (r = 0.78 to 0.90).
12. For numeracy, the primary schools that show more than expected average performance are likely to show more than expected increase in performance over time and vice versa (r = 0.27 to 0.43). However, for literacy, the primary schools that show more than expected average performance are highly likely to show less than expected increase in performance over time and vice versa (r = -0.61 to -0.84).
13. For numeracy, the Type A school effects for each cohort that are estimated using a two-level model consistently agree (r = 0.60 to 0.75) with the overall stable Type A school effects that are estimated using the three-level longitudinal model. However, for literacy Type A school effects for each cohort that are estimated using the two-level model do not always agree (r = 0.47 to 0.77) with the overall stable Type A school effects that are estimated using the longitudinal model.
14. For numeracy as well as literacy, there are small but positive correlations (0.16 to 0.40) between the Type A effects estimated from the two-level models using the four cohorts of students. Thus, a considerable number of schools that show more than expected performance on one testing occasion also show more than expected performance on the other testing occasions and vice versa. However, schools are not highly consistent in terms of the relationship between their performance on one testing occasion and their performance on the other testing occasions and therefore ranking of schools based on data on a single cohort of students could be misleading.
15. Based on the Type A effects for each cohort that are estimated from the two-level model, only a small number of schools that show more than expected performance in numeracy on one testing occasion show more than expected performance in literacy on the other testing occasions (r =0.11 to 0.34). However, within the same testing occasion, a considerable number of schools that show more than expected performance in numeracy also show more than expected performance in literacy (r = 0.54 to 0.64).

Clearly, the most striking finding in this chapter is (12) above. All the attempts made to provide an explanation to (12) above are fruitless and, therefore, it is unclear why the relationship between the stable school effects and the change school effects for numeracy is different from that of literacy (see Appendix 14.6 and summary presented above). However, it must be emphasized that the reliability estimates of the stable school effects and the change school effects are generally found to be low (results in Table 9.1). Bryk and Raudenbush (1992) and Raudenbush (1995) argue that it is difficult to detect systematic relationships between status and growth when the reliability estimates are low. Consequently, the correlation coefficients computed in this chapter provide only a general picture and may not reflect the true extent of the linear relationships between the stable school effects and the change school effects.

Regardless of the so-called 'real world' nature of relationship between the stable school effects and the change school effects, the findings in this chapter have brought to light some information that could form a basis for further research. Apart from (12) above, clearly, there is the need for a further study to investigate why the correlation

between the stable school effects and the change school effects for literacy is positive at Grade 3 and negative at Grade 5. One plausible explanation could lie in the argument provided by Masters and Forster (2000) regarding potential consequences of mastery based large scale testing programs:

> … in high-stakes contexts, teaching and learning can be focused on ensuring that low-achieving students are brought up to the level of the minimum standard. This is a highly desirable outcome for low-achieving students. The implications for students already performing well above the minimum may be less desirable if they are not challenged and extended by classroom teaching and by the assessment themselves. (Masters and Forster, 2000; p. 20)

Hence, it is likely that primary schools in South Australia might be ceasing to provide their Grade 5 students with challenging reading and language experiences once the students have acquired the minimum literacy skills needed to pass the BST. For this argument to hold, it would mean that for some reason, the schools continue to provide their Grade 5 students with challenging experiences in numeracy beyond what is required to pass the BST.

Another plausible explanation is that, unlike numeracy skills, the literacy skills are not purely learned in school. The students could acquire some of the literacy skills outside their schoolwork (for example, at home watching television, reading advertisement, mails and so on). It would appear that students who have limited literacy skills at Grade 3 tend to gain most of the skills required by Grade 5. However, for this argument to hold, it would mean that the control of prior achievement is not enough to remove all the entangled home background effect. Furthermore, it would mean that home background could have some differential effects between the two grades for less able students and more able students. The data available for the current study do not contain sufficient information to establish what is happening.

Potential implications

The results of the longitudinal model HLM analyses agree strongly with what is found in the preceding chapters. That is, after controlling for student-level factors: (a) the amount of variance left unexplained at the school-level is very small (less than three per cent), and (b) the amount of variance left at the student-level is relatively large (about 10 times larger). In other words, more variability between the students is left unexplained and almost all variability between the schools is explained in the longitudinal model.

Despite the improvement possible in the stability of the ranks assigned to schools with the longitudinal design, the variance left unexplained at the school-level is still very small and should raise concern if the model were to be used for ranking purposes. As argued in Chapter 8, it is the amount of variance that is left unexplained at the school-level that is important in the stability of ranks assigned to school based on the so-called 'value added' scores. Although either type of school effect can influence within school variation by interacting with student background (Raudenbush and Willms, 1995), it is the amount of variance left unexplained that is ultimately important in the stability of the ranks assigned to schools based on either type of school effect.

Furthermore, the above findings raise an important question that could have substantial implications for the policy in funding of primary schools. That is, what makes the variance in students' performance in the basic skills of numeracy and literacy within the primary schools in South Australia so large? There are several possible answers to this question.

First, the data available for the current study lack some variables that might explain some of the variability between the students. It is highly likely that inclusion of other variables at the student-level (such as SES[28]) could bring down the amount of variance left unexplained at that level. However, it is unlikely that inclusion of a SES variable would drop the variance left unexplained substantially. This is because, in South Australia, it is reasonable to expect that most of the variance associated with the SES is entangled with other student background variables such as Prior Achievement and Racial Background and, therefore, has already been catered for in the model.

Second, measurement errors in the outcome variable could be a source of the observed variability between students. For example, in Rasch scaling the abilities of the students at the extremes (near perfect or near zero raw scores) are estimated with larger errors compared to the abilities of the students with average (or near average) raw scores. Furthermore, equating across the occasions might have introduced errors that could have inflated (or deflated) scores for certain occasions thus causing misleading variability between students.

Finally, there is the possibility that the tests used are unreliable for some ability categories of the students. If this is the case, then certain students might have attained scores that do not reflect their actual ability level, and this might be the cause of the variability observed between the students.

Whatever the cause of the variability between students, it is obvious that this variability is substantially large compared to the variability between schools. However, it should be emphasized that, although the variability between schools is small, this does not imply that it is insignificant. Indeed, there are no research-backed limits as to how small the variability should be for the variability to be considered trivial. The argument here is entirely on the stability (and consequently, the usefulness) of the ranks assigned to schools based on small amounts of variance left unexplained rather than the significance of the variance left unexplained at the school-level for either type of school effects.

Plainly, it is difficult to identify reliably weak schools. Thus, it appears that the practice of identifying weak schools and providing them with funds would not seem appropriate. It appears that the government should focus on identifying weak students within schools and providing them with remedial programs to help them gain the required skills.

The next chapter focuses on the estimation of school effects where allowances are made for student background characteristics, school context and school characteristics.

[28] Socioeconomic status

10

Type C School Effects

In the computation of Type B effects, it is generally accepted that only the student-related school-level variables are included in the model so that the amount of between-school variance left can be attributed directly to the schools as such rather than the students who attend them. Consequently, the student free school-level variables are excluded in order to estimate how much of the variance is taken up by the characteristics of the student population in the school (Harker and Nash, 1996).

However, for primary schools in South Australia, it might be appropriate to include in the model some of the school characteristic variables, especially School Size (SSIZE or SSIZE_2) and School Location (METRO). This is because school size and school locality (rural/urban) are among the characteristics of the school that are taken into account by the government when providing funds to the primary schools.

This leads to another type of school effect (Type C), which represents the increment to student achievement attributed to school *j* after controlling for the effects of student intake and the effects of the average school context and school characteristics. In this case, the residual term ($\mathbf{u}_{00j}$) in the longitudinal design includes mostly effects attributed to school policy and, therefore, it is the stable Type C effect. In this case, the 'school-by-occasion' interaction effect (that is, $\mathbf{u}_{01j}\mathbf{OCC}_{jk}$ component) in the longitudinal structure shows how the school performance has changed over time. And the residual term ($\mathbf{u}_{00j}$) includes mostly those effects attributed to school policy and, therefore, it is the stable Type C effect. It is this stable Type C effect component that would be of interest because it indicates the curricular, instructional and managerial effects of the school. In particular, the stable Type C effect involves:

(a) quality of teaching in the school;

(b) curriculum planning and implementation in the school;

(c) management of the school, involving effective utilization of time; and

(d) level of student motivation and perseverance that is independent of school context.

However, in South Australia, because only small amounts of variance are left unexplained at the school-level in the Type B models described in Chapter 9, it is

expected that the amount of variance left at the school-level in Type C effects model would generally be very small. Therefore, for purposes of ranking schools in South Australia, it is questionable to compute this type of effect. Nevertheless, the Type C effects are computed here because it is considered that policy makers and school officials might wish to have a general picture of these indices since they reflect the curricular, instructional and managerial effects of the school.

The outline of this chapter is as follows. The first two sections describe the specific model for the estimation of the Type C school effects based on the longitudinal structure and the estimation of these indices of school effectiveness, respectively. The third section presents the results of the analyses with comparisons being made between the results obtained here from the Type C effects models and the results obtained in Chapter 9 from the Type B effects models. The final section focuses on the correlations between (a) Type C effects, and (b) Type C effects and Type B effects computed in Chapter 9.

All the multilevel analyses reported in this chapter are carried out using the HLM5/3L computer program (Raudenbush, Bryk and Congdon, 2000).

Specification of Type C effects model

In the longitudinal modelling of the Type C effects, allowance is made for student background characteristics, school context and school characteristics (size and location). At the micro-level, the model for estimation of Type C effects is exactly the same as the general longitudinal models for estimation of Type A and Type B effects described in Chapter 9. However, at the meso- and macro-levels of the Type C effect models, allowance is made for school context and school characteristics (size and location). Because school size is a "malleable" characteristic (see Postlethwaite and Ross, 1992; p.3), the variable SSIZE (School Size) can be examined for possible inclusion at the meso-level of the model and the variable SSIZE_2 (Average School Size over the study period) can be examined for possible inclusion at the macro-level of the model. The variable METRO (School Location, urban or rural) can be examined for possible inclusion at the macro-level of the model because it is a stable characteristic. In the analyses carried out in this chapter, the raw school size variables (SSIZE, SSIZE_2) are preferred to the log-transformed versions of the variables (SSIZELOG, SSIZEL_2) to make it easier to interpret the results and also in order to cater for the amount of variance attributed directly to the actual size of the school.

Following the notations and arguments introduced in Chapter 9, the three-level longitudinal model for the estimation of Type C effects, can be described as follows:

Level-1 model

$$\mathbf{Y}_{itj} = \boldsymbol{\pi}_{0tj} + \boldsymbol{\pi}_{htj}\boldsymbol{X}_{hitj} + \mathbf{e}_{itj} \qquad \text{Equation 10.1}$$

Level-2 model

$$\boldsymbol{\pi}_{0tj} = \boldsymbol{\beta}_{00j} + \boldsymbol{\beta}_{01j}\mathbf{OCC}_{tj} + \boldsymbol{\beta}_{02j}\mathbf{SSIZE}_{tj} + \boldsymbol{\beta}_{0gj}\bar{\boldsymbol{X}}_{gtj} + \mathbf{r}_{0tj} \qquad \text{Equation 10.2}$$

$$\boldsymbol{\pi}_{htj} = \boldsymbol{\beta}_{h0j} \qquad \text{Equation 10.3}$$

Level-3 model

$$\boldsymbol{\beta}_{00j} = \boldsymbol{\gamma}_{000} + \boldsymbol{\gamma}_{001}\boldsymbol{METRO}_j + \boldsymbol{\gamma}_{002}\boldsymbol{SSIZE\_2}_j + \boldsymbol{\gamma}_{00f}\acute{\boldsymbol{x}}_{00fj} + \mathbf{u}_{00j} \qquad \text{Equation 10.4}$$

$$\boldsymbol{\beta}_{01j} = \boldsymbol{\gamma}_{010} + \boldsymbol{\gamma}_{011}\boldsymbol{METRO}_j + \boldsymbol{\gamma}_{012}\boldsymbol{SSIZE\_2}_j + \boldsymbol{\gamma}_{01f}\acute{\boldsymbol{x}}_{01fj} + \mathbf{u}_{01j} \qquad \text{Equation 10.5}$$

$$\boldsymbol{\beta}_{02j} = \boldsymbol{\gamma}_{020} \qquad \text{Equation 10.6}$$

$\beta_{0gj} = \gamma_{0g0}$ Equation 10.7

$\beta_{h0j} = \gamma_{h00}$ Equation 10.8

where:

β_{02j} is the regression coefficient associated with SSIZE for school *j*;

γ_{001} is the regression coefficient associated with METRO;

γ_{002} is the regression coefficient associated with SSIZE_2;

γ_{011} is the regression coefficient associated with the interaction effect between OCC and METRO; and

γ_{012} is the regression coefficient associated with the interaction effect between OCC and SSIZE_2;

All the other components in Equations 10.1 to 10.8 carry the same meaning as described in Chapter 9 for Type B effects model. In the HLM analyses, the predictor OCC is group-mean centred, therefore, β_{00j} is the mean effectiveness of school *j* during the period of the study (Kreft et al., 1995) and β_{01j} is the difference in school *j* trend in achievement relative to the overall trend (Willms and Raudenbush, 1989).

In the above model, for purposes of simplicity, the terms $\boldsymbol{X}$, $\bar{\boldsymbol{X}}$ and $\acute{\boldsymbol{x}}$ are used to represent several independent variables that describe student characteristics, several variables that describe the school context at each testing occasion, and several variables that describe the average school context over the study period, respectively. Again for purposes of simplicity, the effects associated with the student background characteristics, π_{htj} (Equation 10.3), β_{h0j} (Equation 10.8), the effects associated with changing school context, β_{0gj} (Equation 10.7), and the effects associated with changing school size β_{02j} (Equation 10.6) are specified as fixed. However, in the actual analyses, these effects are only specified as fixed if they do not vary significantly across the occasions or across the schools.

Employing the same procedure followed to develop a single linear equation for the estimation of Type B effects, Equations 10.1, 10.2, 10.4 and 10.5 can be combined to form the following single equation.

$$
\begin{aligned}
\mathbf{Y}_{itj} &= [\gamma_{000}] && \text{(grand mean)}\\
&+ [\gamma_{010}\mathbf{OCC}_{tj}] && \text{(main effect of occasion)}\\
&+ [\pi_{htj}\boldsymbol{X}_{hitj}] && \text{(control for student intake)}\\
&+ [\gamma_{001}\boldsymbol{METRO}_j + \gamma_{002}\boldsymbol{SSIZE\_2}_j \\
&\quad + \gamma_{00f}\acute{\boldsymbol{x}}_{00fj} + \mathbf{u}_{00j}] && \text{(stable component of school effect)}\\
&+ [\gamma_{011}\boldsymbol{METRO}_j * \mathbf{OCC}_{tj} + \gamma_{012}\boldsymbol{SSIZE\_2}_j * \mathbf{OCC}_{tj} + \\
&\quad \gamma_{01f}\acute{\boldsymbol{x}}_{01fj}\mathbf{OCC}_{tj} + \mathbf{u}_{01j}\mathbf{OCC}_{tj} + \beta_{02j}\mathbf{SSIZE}_{tj} + \\
&\quad \beta_{0gj}\bar{\boldsymbol{X}}_{gtj} + \mathbf{r}_{0tj}] && \text{(change component of school effect)}\\
&+ [\mathbf{e}_{itj}] && \text{(student-level error)}
\end{aligned}
$$

Equation 10.9

In the above model, the stable component of school effect now has four terms: $\gamma_{001}\boldsymbol{METRO}_j$, $\gamma_{002}\boldsymbol{SSIZE\_2}_j$, $\gamma_{00f}\acute{\boldsymbol{x}}_{00fj}$, and $\mathbf{u}_{00j}$. The term $\gamma_{001}\boldsymbol{METRO}_j$ represents control for school location while the terms $\gamma_{002}\boldsymbol{SSIZE\_2}_j$ and $\gamma_{00f}\acute{\boldsymbol{x}}_{00fj}$ represent the control for average school size and the average school context over the duration of the study, respectively. The residual term $\mathbf{u}_{00j}$ now represents the increment to student

achievement attributable to school *j* after controlling for the effects of student intake and the effects of the average school context and school characteristics (location and size). Hence, this residual term ($\mathbf{u}_{00j}$) includes mostly those effects attributed to school policy and, therefore, it is the stable Type C effect.

In Equation 10.9, the change component of school effect has seven terms:

(i) $\gamma_{011}METRO_j * \mathbf{OCC}_{tj}$, a 'locality-by-occasion' interaction effect;

(ii) $\gamma_{012}SSIZE\_2_j * \mathbf{OCC}_{tj}$, a 'size-by-occasion' interaction effect;

(iii) $\gamma_{01f}\acute{x}_{01fj}\mathbf{OCC}_{tj}$, a 'context-by-occasion' interaction effect;

(iv) $\mathbf{u}_{01j}\mathbf{OCC}_{tj}$, a 'school-by-occasion' interaction effect;

(v) $\beta_{02j}\mathbf{SSIZE}_{tj}$, control for changing school size;

(vi) $\gamma_{0g0}\bar{X}_{gtj}$, control for changing school context; and

(vii) $\mathbf{r}_{0tj}$, a random year-to-year fluctuation in a school's intake-adjusted levels of performance.

Here again the main interest is in concern for the value of the 'school-by-occasion' interaction effect ($\mathbf{u}_{01j}\mathbf{OCC}_{tj}$), because it represents the systematic change in the performance of the school after allowance has been made for student characteristics, school context and school characteristics. That is, $\mathbf{u}_{01j}\mathbf{OCC}_{tj}$ term shows the change that has occurred in a school's Type C effect over the study period.

Estimation of Type C effects

The model for Type C effects specified above (Equation 10.9) is estimated for numeracy and literacy using the two data sets. Basically, the same procedure followed for the estimation of the Type B effects in Chapter 9 is followed here for the estimation of the Type C effects. However, unlike in the estimation of Type B effects where the student free school-level variables were excluded from the analyses, here the school size variables (SSIZE and SSIZE_2) and the locality of the school variable (METRO) are included in the analyses. The school context variables as well as the school size and locality of the school variables are included in the examination of possible cross-level interaction effects.

As mentioned above, school size and locality of the school are among the factors that are taken into consideration by the government when allocating funds to the primary schools in South Australia. Thus, a decision is made here to include in the model the variables SSIZE_2 and METRO regardless of the statistical significance of the effects of these variables. In addition, it is considered that school officials and policy makers might wish to adjust for the effects of Proportion of School Cardholders, Absenteeism and Mobility Rates in the school even if these effects of school context do not meet the $p<0.05$ criterion of statistical significance. Consequently, a decision is made to include in the model the variables PSCARD_2, ABSENT_2 and MOBILI_2 even if the effects of these variables are not statistically significant at the $p<0.05$ level.

For this study, however, the HLM estimation procedure fails to proceed when the mean performance level of each school (β_{00j} in Equation 10.4) and when the variation between schools in their trend component (β_{01j} in Equation 10.5) are simultaneously regressed on all the five variables. Consequently, a decision is made here to regress the mean performance of each school on all the five variables whether or not the effects of the variable is statistically significant, and to regress the variation between schools in their trend component on those variables that show statistically significant

($p<0.05$) interaction effects with the change slope. Apart from the five variables (SSIZE_2, METRO, PSCARD_2, ABSENT_2 and MOBILI_2) all the other variables are included in the intercept model (Equation 10.4) only if they meet the $p<0.05$ criterion of statistical significance.

At the micro-level and the meso-level, the models for the estimation of Type C effects are exactly the same as the corresponding models for estimation of Type B effects described in Chapter 9.

Results

The results of HLM analyses described above for the estimation of Type C effects provide estimates of reliability, fixed effects, variance components and the deviance statistics. These results are discussed in separate sub-sections below.

Reliability estimates

Table 10.1 displays the estimated reliabilities of the stable and the change components of the Type C effects at the third level of models. The reliability estimates of all the variables with random effects at the second and the third levels of the Type C effect models can be found in Appendix 14.3.

For the same subject area, the results in Table 10.1 show that the reliability estimates of the stable and change Type C effects observed using the transience data set follow closely those obtained using the non-transience data set, which is consistent with what is found in Chapter 9 for Type A and Type B effects. The results in Table 10.1 also show that the stable effects (λ = 0.366 to 0.442) are estimated more reliably than change effects (λ = 0.133 to 0.162), and that in general, the reliability estimates for numeracy are slightly higher than for literacy. Again, these results are consistent with what is found for Type A and Type B effects.

Finally, it should be noted that the reliability estimates from the Type C effects models presented in Table 10.1 follow closely the reliability estimates from the corresponding Type B effects models presented in Chapter 9. For example, for literacy and using the transience data set, the reliability estimate from the Type B effects model for the stable components is 0.403, which follows closely the corresponding estimate from the Type C effects model, 0.402. Nonetheless, it should be noted that all the reliability estimates for the stable and the change components obtained from the Type C effects models are marginally lower than the corresponding estimates obtained from the Type B effects models. Hence, it could be argued that, Type C effects are estimated a little bit less reliable than are Type B effects.

Table 10.1 School-level reliability estimates from the Type C effects models

		Numeracy		Literacy	
		Tran$^{\propto}$	Non-Tran$^{\beta}$	Tran$^{\propto}$	Non-Tran$^{\beta}$
Effect	*Random level-2 coefficient*				
Stable	**INTRCPT1**/ INTRCPT2,	0.433	0.442	0.402	0.366
Change	**INTRCPT1**/ OCC,	0.162	0.153	0.153	0.133

Notes: $\propto$ - Using all the students matched (Schools = 482).
β - Using only those students matched in the same school (Schools = 479).

Deviance statistics

In Table 10.2, the deviance statistics and chi-square test are used to compare the fit of the Type C effects model to the fit of the corresponding Type B effects model from Chapter 9. The results in Table 10.2 show a significant drop in deviance statistic in the Type C effect models as indicated by the p-value ($p<0.05$) of the chi-square test for each pair of models compared. Therefore, it can be concluded that the inclusion of the school characteristics as predictors in the models significantly improves the overall fit of the models.

Table 10.2 Comparison of model fit using chi-square tests

		Deviance Statistic	Number of Parameters	Chi-square Statistic	Degrees of Freedom	P-value
Numeracy						
Tran$^{\propto}$	Type B	87773.15	31	340.06	9	0.00
	Type C	87761.37	35	11.78	4	0.02
Non-Tran$^{\beta}$	Type B	75168.28	19	183.20	2	0.00
	Type C	75158.82	23	9.46	4	0.05
Literacy						
Tran$^{\propto}$	Type B	79776.80	27	141.52	1	0.00
	Type C	79765.30	30	11.50	3	0.00
Non-Tran$^{\beta}$	Type B	67861.54	24	82.59	3	0.00
	Type C	67845.90	28	15.64	4	0.00

Notes: $\propto$ - Using all the students matched (Schools = 482).
β - Using only those students matched in the same school (Schools = 479).

Fixed effects

The estimations of the fixed effects for the models for Type C effects are presented in Table 10.3. At the micro-level and the meso-level, the results displayed in Table 10.3 follow closely the corresponding results of fixed effects from the Type B effects models presented in Chapter 9 and, therefore, no additional discussions of the results for these two levels are necessary here.

At the macro-level, the results for Type C effects (Table 10.3) show that two variables, namely PSCARD_2 (Average Proportion of School Cardholders), and ABSENT_2 (Average Absenteeism Rate) have significant influences on achievement in numeracy and literacy regardless of the data set used. Based on a $p<0.05$ criterion, the results in Table 10.3 show that although the Average Mobility Rate (MOBILI_2) and the Average School Size (SSIZE_2) variables are included in the models for the estimation of Type C effects, they have no significant effects on any of the two outcome measures.

In addition, the results in Table 10.3 show that the variable METRO (School Location; coded urban=1, rural=0), has a significant ($p<0.05$) influence on literacy but not on numeracy. These results seem to be contrary to those of the analyses presented in Chapters 7 and 8, which indicate that the mobility rate of a school, size of a school and locality of a school have significant influences on achievement in numeracy and literacy. But again it should be borne in mind that the models specified in this chapter and Chapter 9 are different from the models specified in Chapters 7 and 8.

At the macro-level, considering the transience data set, the results in Table 10.3 show that Average Prior Achievement (Y3NSCO_2 or Y3LSCO_2) is positively related to

each of the two outcome measures, which is consistent with what is found from Type B effects models. However, unlike in Type B effects model, the results here indicate that the variable INOZ_2 (overall Average Living in Australia) has no significant effect on achievement after the school characteristics and the mobility rate of the school variables are included in the model.

For numeracy, when considering the transience data set, the results in Table 10.3 indicate cross-level interaction effects between these variables:

(a) Prior Achievement (Y3NSCORE) and Average Living in Australia (INOZ_1);

(b) Sex of the Student (SEX) and overall Average Prior Achievement (Y3NSCO_2);

(c) Transience (TRANS) and overall Average Age of the Students (AGE_2); and

(d) Transience (TRANS) and overall Average Prior Achievement (Y3NSCO_2).

Generally, the above cross-level interaction effects are the same as the interaction effects outlined in Chapter 7, and therefore, it is considered unnecessary to discuss these interaction effects again here.

For both numeracy and literacy, and for both data sets, the results in Table 10.3 also show significant interaction effects between OCC and SSIZE_2. Figures 10.1 and 10.2 show the graphical representations of the interaction effects between OCC with SSIZE_2 and the outcome variable Y5NSCORE or Y5LSCORE for numeracy and literacy, respectively. The coordinates of the graphs are calculated from the final estimation of the fixed effects obtained from the Type C effects models using all the students who could be matched (results in Table 10.3). The procedure described by Lietz (1996) for the calculation of the coordinates is followed here. A detailed account of this procedure can also be found in Hungi (2003; pp.503-528).

The graphical representation in Figures 10.1 and 10.2 shows that during the earlier testing occasions, students in schools with many pupils are generally estimated to achieve better in numeracy and literacy than their counterparts in small schools. However, during the later testing occasions, students in small schools are estimated to achieve better in both subjects than students in large schools.

Figure 10.1 also shows that the students in small schools are estimated to perform equally well in numeracy regardless of the testing occasions while students in large schools are estimated to perform lower on the later testing occasion than on the earlier testing occasion. However, Figure 10.2 shows that, regardless of the size of the school, students are generally estimated to perform lower in literacy on the later testing occasion than on the earlier testing occasion. Moreover, the decline over time is greater for larger schools than for smaller schools.

Stable and change variance components

Table 10.4 presents the values of the variance components, the degrees of freedom, the chi-square statistics and the p-value associated with the stable and the change components of school effects for the two outcome measures, using the two data sets. The variance components associated with all the variables with random effects at the second and the third levels of the Type C effects models can be found in Appendices 14.4 and 14.5.

Clearly, the results in Table 10.4 show that there is significant variability in the primary schools in South Australia in terms of the stable and change Type C school effects for numeracy and literacy.

Table 10.3 Final estimation of fixed effects from Type C effects models

1) For numeracy

			Transience Data Set (Schools = 482)					**Non-Transience Data Set** (Schools = 479)				
			Std'zed	Metric	SE	T-ratio	P-value	Std'zed	Metric	SE	T-ratio	P-value
INTRCPT1,												
INTRCPT2,	INTRCPT3	G000	1.34	1.34	0.01	95.67	0.00 *kj*	1.39	1.38	0.02	91.37	0.00 *kj*
	Y3NSCO_2	G001	0.07	0.11	0.03	3.86	0.00	×××	×××	×××	×××	×××
	ABSENT_2	G002	-0.11	-2.11	0.53	-4.00	0.00	-0.10	-1.97	0.50	-3.96	0.00
	PSCARD_2	G003	-0.07	-0.38	0.09	-4.48	0.00	-0.10	-0.57	0.08	-7.26	0.00
	MOBILI_2	G004	-0.04	-0.00	0.00	-1.71	0.09 *ξ*	-0.03	-0.00	0.00	-1.41	0.16 *ξ*
	SSIZE_2	G005	0.01	0.00	0.00	0.64	0.52 *ξ*	-0.01	0.00	0.00	-0.82	0.41 *ξ*
	METRO	G006	-0.00	-0.00	0.02	-0.09	0.93 *ξ*	0.01	0.02	0.02	0.92	0.36 *ξ*
OCC,	INTRCPT3	G010	-0.02	-0.02	0.01	-2.37	0.02 *j*	-0.02	-0.02	0.01	-2.07	0.04 *j*
	SSIZE_2	G011	-0.02	0.00	0.00	-2.89	0.00	-0.02	0.00	0.00	-2.93	0.00
AGE_1		B02	-0.03	-0.20	0.08	-2.54	0.01	×××	×××	×××	×××	×××
HOME_1		B03	0.05	0.17	0.07	2.44	0.02	×××	×××	×××	×××	×××
SEX,	INTRCPT3	G100	-0.04	-0.09	0.01	-10.09	0.00 *j*	-0.05	-0.10	0.01	-10.67	0.00 *j*
	Y3NSCO_2	G101	-0.01	-0.05	0.02	-2.02	0.04	×××	×××	×××	×××	×××
AGE,		B20	-0.07	-0.19	0.01	-15.59	0.00	-0.07	-0.19	0.01	-14.37	0.00
ATSI,		B30	0.03	0.19	0.02	7.82	0.00	0.03	0.19	0.03	7.08	0.00
INOZ,		B40	-0.01	-0.04	0.01	-2.82	0.01	-0.01	-0.04	0.01	-2.88	0.00
TRANS,	INTRCPT3	G500	-0.04	-0.11	0.01	-8.30	0.00 *j*					
	AGE_2	G501	-0.02	-0.49	0.24	-2.09	0.04					
	Y3NSCO_2	G502	0.03	0.13	0.03	4.38	0.00					
Y3NSCORE,												
INTRCPT2		B60	0.69	0.55	0.01	115.43	0.00 *k*	0.70	0.57	0.01	116.72	0.00 *k*
INOZ_1		B61	0.01	0.11	0.05	2.31	0.02	×××	×××	×××	×××	×××

(Continued)

Table 10.3 Final estimation of fixed effects from Type C effects models *(Continued)*

2) For literacy

			Transience Data Set (Schools = 482)					**Non-Transience Data Set** (Schools = 479)				
			Std'zed	Metric	SE	T-ratio	P-value	Std'zed	Metric	SE	T-ratio	P-value
INTRCPT1,												
INTRCPT2,	INTRCPT3	G000	1.45	1.44	0.01	108.61	0.00 *kj*	1.50	1.48	0.01	112.64	0.00 *kj*
	Y3LSCO_2	G001	0.04	0.07	0.03	2.54	0.01	×××	×××	×××	×××	×××
	ABSENT_2	G002	-0.08	-1.63	0.65	-2.50	0.01	-0.08	-1.59	0.68	-2.35	0.02
	PSCARD_2	G003	-0.05	-0.26	0.08	-3.26	0.00	-0.06	-0.36	0.07	-4.85	0.00
	MOBILI_2	G004	-0.04	-0.00	0.00	-1.82	0.07 ξ	-0.03	-0.00	0.00	-1.58	0.11 ξ
	SSIZE_2	G005	0.01	0.00	0.00	0.53	0.60 ξ	0.00	0.00	0.00	0.02	0.99 ξ
	METRO	G006	0.02	0.04	0.02	2.02	0.04	0.02	0.04	0.02	2.48	0.01
OCC,	INTRCPT3	G010	-0.10	-0.09	0.01	-13.89	0.00 *j*	-0.10	-0.09	0.01	-13.20	0.00 *j*
	SSIZE_2	G011	-0.02	0.00	0.00	-2.85	0.01	-0.02	0.00	0.00	-2.87	0.01
SEX,		B10	0.02	0.04	0.01	5.27	0.00 *j*	0.02	0.03	0.01	4.00	0.00 *j*
AGE,		B20	-0.06	-0.18	0.01	-16.45	0.00	-0.06	-0.17	0.01	-15.25	0.00
ATSI,		B30	0.03	0.15	0.02	6.63	0.00	0.03	0.14	0.03	5.35	0.00
HOME,		B40	0.02	0.03	0.01	3.32	0.00	0.01	0.02	0.01	2.74	0.01
INOZ,		B50	-0.02	-0.05	0.01	-4.25	0.00	-0.02	-0.05	0.01	-4.39	0.00
TRANS,		B60	-0.02	-0.06	0.01	-4.62	0.00					
Y3LSCORE,		B70	0.77	0.61	0.00	142.80	0.00 *kj*	0.78	0.62	0.00	139.62	0.00 *kj*

Notes: Shade - The variable TRANS is not available for examination in this model.
××× - Variable has no significant effect and, therefore, excluded in this model.
ξ - Variable has no significant effect (p>0.05) but included in the model.
k - Residual parameter of this coefficient is left to vary at the occasion-level
j - Residual parameter of this coefficient is left to vary at the school-level
- The standard errors (SE), t-ratios and p-values presented are those obtained using unstandardized variables.

Std'zed - Regression coefficient obtained using standardized variables.
Metric - Regression coefficient obtained using unstandardized variables.

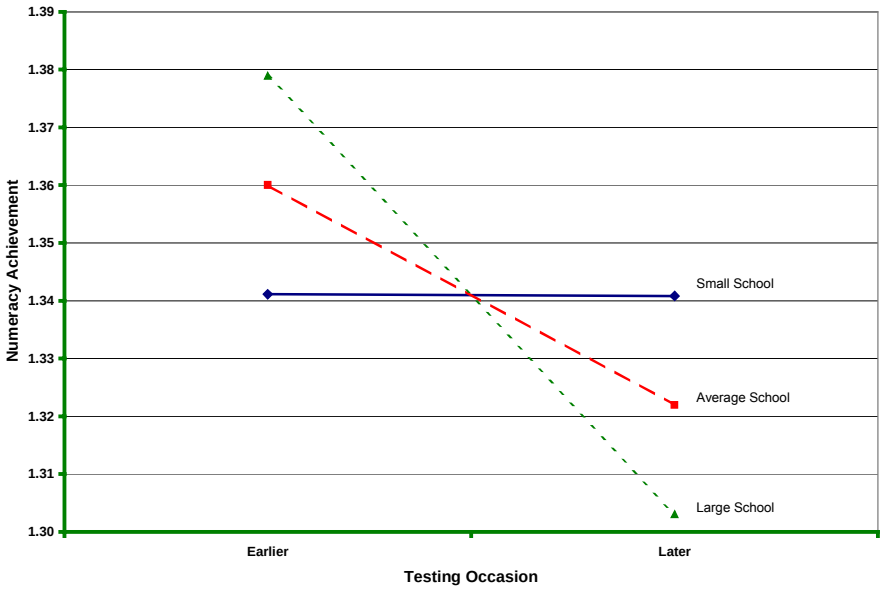


Figure 10.1 Impact of the interaction effect of School Size with testing Occasion on Numeracy performance

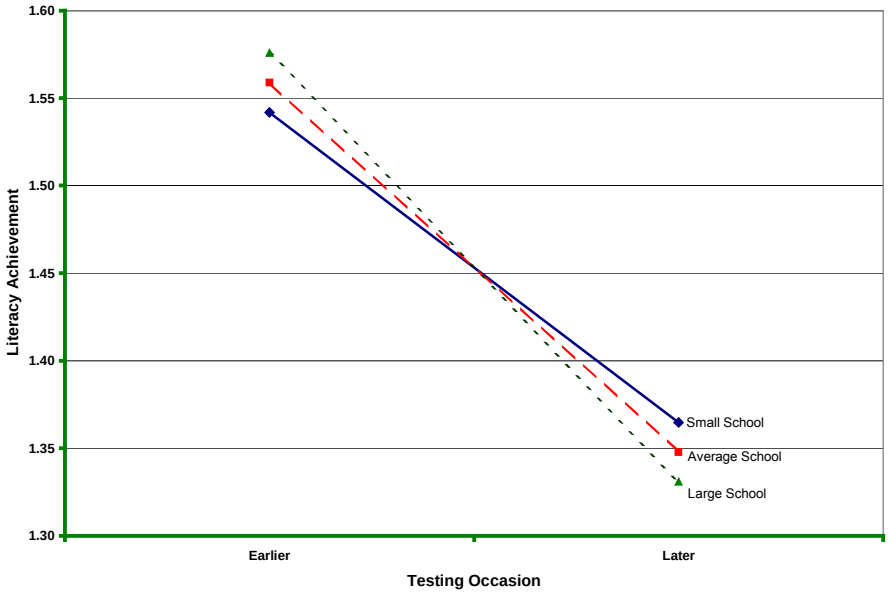


Figure 10.2 Impact of the interaction effect of School Size with testing Occasion on Literacy performance

The stability ratios (in **bold** in Table 10.4) are for the Type C effects estimated in this chapter and the figures in *italics* (immediately below the figures in bold in Table 10.4) are the stability ratios for the corresponding Type B effects estimated in Chapter 9.

Arguably, based on the stability ratio criterion (Willms and Raudenbush, 1989) the Type C effects are marginally more stable than the Type B effects. However, it appears that the stability of Type C effects do not differ considerably when compared across the two outcome measures and across the two data sets, which is consistent with what is found for Type A and Type B effects in Chapter 9. Nevertheless, as in the case of Type A and Type B effects, the stability ratios for numeracy are in general slightly higher than the stability ratios for literacy, which seems to emphasize that the school effects on numeracy are to some extent more stable than school effects on literacy. This result supports the view that schools change with respect to teaching of literacy more than they change with respect to teaching of numeracy.

Table 10.4 Final estimation of variance components from Type C effects models

		Transience Data Set (Schools = 482)					**Non-Transience Data Set** (Schools = 479)				
		Var. Comp.	df	Chi-Square	P-value	Stability Ratio	Var. Comp.	df	Chi-Square	P-value	Stability Ratio
Numeracy											
	Stable (u_{00j})	0.017	456	841.379	0.000	**5.42** *4.94*	0.020	457	862.577	0.000	**5.84** *5.17*
	Change (u_{01j})	0.003	461	584.793	0.000		0.003	461	586.890	0.000	
Literacy											
	Stable (u_{00j})	0.014	460	817.544	0.000	**5.12** *4.79*	0.013	457	737.365	0.000	**5.13** *4.73*
	Change (u_{01j})	0.003	465	579.336	0.000		0.003	461	565.698	0.001	

Variance partitioning and variance explained

Table 10.5 gives comparisons of the amounts of variances left unexplained at the student-level, occasion-level, school-level and in all the levels combined that were estimated using the Type B effects models, and those that were estimated using the Type C effects models for numeracy and literacy. The information in Table 10.5 shows that the amounts of variance left unexplained in the Type B effects models and the amounts of variance left unexplained in the Type C effects models are basically the same regardless of the data set used. Obviously, the inclusion of the school characteristics in the Type C effects models offered no added advantage as far as the amounts of variance explained in the final models are concerned.

As in Type B effects models, the results in Table 10.5 show that the percentages of total variance left unexplained at the school-level (Level-3) are small (about one per cent) in Type C effects models as well. These results are consistent across the two subjects (numeracy and literacy) and across the two data sets used. However, when interpreting these results, it should be remembered that the variation attributed to school effects could be larger than the overall variation between schools because school effects can also influence within school variation by interacting with student background characteristics (see Raudenbush and Willms, 1995; p.316).

Table 10.5 Percentages of variance left unexplained in Type B and Type C effects models

		Tran$^{\propto}$				Non-Tran$^{\beta}$			
		Level-1	Level-2	Level-3	Total	Level-1	Level-2	Level-3	Total
Numeracy									
	Var. Avail. (%)	65.0	19.1	16.0		68.8	16.0	15.2	
Type B	Var. Left (%)	36.2	2.3	1.1	39.7	37.6	2.9	1.4	41.9
Type C	Var. Left (%)	36.2	2.3	1.1	39.7	37.6	2.8	1.4	41.8
Literacy									
	Var. Avail. (%)	69.3	15.8	15.0		71.3	15.2	13.5	
Type B	Var. Left (%)	31.2	2.4	1.0	34.5	31.5	2.7	1.0	35.2
Type C	Var. Left (%)	31.2	2.4	1.0	34.5	31.5	2.7	1.0	35.2

Notes: $\propto$ - Using all the students matched (Schools =482).
β - Using only those students matched in the same school (Schools = 479).

Correlations

The first part of this section focuses on the correlations between the indices of individual school effects computed in the current chapter and the second part focuses on the correlations between the Type C indices and the Type B indices computed in Chapter 9. For presentation purposes, the same system used in naming of school effects in the previous chapter is used in this section. Hence, the codes that start with a 'T' are for school effects obtained using the transience data set while the codes that start with a 'V' are those for school effects obtained using the non-transience data set. The prefixes 'EB00' and 'EB01' represent the empirical Bayes estimates for the stable and the change components of school effects respectively. An 'NC' ending indicates that the code represent Type C effects for numeracy while an 'LC' indicates that the code represent Type C effects for literacy.

Correlations between Type C school effects

The top panel of Table 10.6 shows the correlations between the stable Type C effects obtained using the transience data set and those obtained using the non-transience data set while the bottom panel of the table shows the corresponding information for change Type C effects.

The figures given in bold in Table 10.6 are the coefficients of the correlations between the stable (or between the change) Type C effects obtained using the transience data set and the ones obtained using the non-transience data set within the same subject area. Based on Type C effects (either stable or change), the results in Table 10.6 show that ranking of schools obtained using all students who could be matched and the ranking obtained using the students who remained in the same school are essentially the same (r = 0.94 to 1.00).

In addition, the results in Table 10.6 show medium to large correlations (0.49 to 0.55) for stable Type C effects and small correlations (0.18 to 0.23) for change Type C effects, across the two subjects regardless of the data set used. The medium to large correlations across the two subjects for stable Type C effects indicate that after adjusting for student intake, school context and school characteristics, a considerable

number of schools that show more than expected performance in numeracy also show more than expected performance in literacy and vice versa.

Table 10.6 Correlations between Type C effects across data sets and across outcome measures

	Numeracy		Literacy	
	Tran$^{\propto}$	**Non-Tran$^{\beta}$**	**Tran$^{\propto}$**	**Non-Tran$^{\beta}$**
Stable**				
	TEB00NC	VEB00NC	TEB00LC	VEB00LC
TEB00NC	1.00			
VEB00NC	**0.98**	1.00		
TEB00LC	0.49	0.49	1.00	
VEB00LC	0.53	0.55	**0.95**	1.00
Change**				
	TEB01NC	VEB01NC	TEB01LC	VEB01LC
TEB01NC	1.00			
VEB01NC	**0.95**	1.00		
TEB01LC	0.23	0.19	1.00	
VEB01LC	0.18	0.18	**0.94**	1.00

Notes: $\propto$ - Using all the students matched (Schools =482).
β - Using only those students matched in the same school (Schools = 479).
** - All the correlations are significant at the 0.01 level.

However, the small correlations for the change Type C effects show that only a small number of schools that record more than expected change in performance in numeracy also record more than expected change in performance in literacy and vice versa.

Table 10.7 displays the correlation between the stable Type C effects and the change Type C effects using the two data sets for each of the outcome measures. The numbers given in bold in Table 10.7 are the correlations between stable effects and change effects within the same subject and for one data set.

For numeracy, the results in Table 10.7 show near large to large but positive correlations (0.48 to 0.50) between the stable effects and the change effects for Type C effects. Hence, based on Type C measures of school effectiveness, a considerable number of schools that record more than expected average performance in numeracy also record more than expected change in performance in numeracy over time and vice versa. This finding is consistent with what is found in the previous chapter for Type B effects.

For literacy, the results in Table 10.7 show very strong to extremely strong negative correlations (-0.71 to -0.80) between the stable school effects and the change school effects for Type C, which is consistent with what is found in the previous chapter for Type A and Type B effects. Thus, based on Type C effects, most schools that record more than expected average performance in literacy record less than expected increase in performance in literacy over time, and vice versa.

Correlations between Type B and Type C school effects

The top panel of Table 10.8 displays the correlations between Type C (both stable and change) effects for numeracy computed in this chapter and the Type B effects (both stable and change) for numeracy computed in Chapter 9, while the bottom panel of the

table displays the corresponding information for literacy. Each panel of the table displays the correlations obtained using the transience data set as well as those obtained using the non-transience data set. The numbers given in bold in the table are the correlations between the stable (or between the change) Type C effects and the stable (or change) Type B effects within the same subject and for the same data set.

Table 10.7 Correlations between stable and change Type C effects

		Change effects	
		Tran$^{\propto}$	**Non-Tran$^{\beta}$**
Stable effects	**Numeracy****		
		TEB01NC	VEB01NC
	TEB00NC	**0.50**	0.48
	VEB00NC	0.49	**0.49**
	Literacy**		
		TEB01LC	VEB01LC
	TEB00LC	**-0.78**	-0.78
	VEB00LC	-0.71	**-0.80**

Notes: $\propto$ - Using all the students matched (Schools =482).
β - Using only those students matched in the same school (Schools = 479).
** - All the correlations are significant at the 0.01 level.

Table 10.8 Correlations between Type B and Type C school effects

		Type B effects			
		Stable		**Change**	
		Tran$^{\propto}$	Non-Tran$^{\beta}$	Tran$^{\propto}$	Non-Tran$^{\beta}$
Type C effects	**Numeracy****				
		TEB00NB	VEB00NB	TEB01NB	VEB01NB
	TEB00NC	**1.00**	0.98		
	VEB00NC	0.98	**1.00**		
	TEB01NC			**0.99**	0.94
	VEB01NC			0.94	**0.99**
	Literacy**				
		TEB00LB	VEB00LB	TEB01LB	VEB01LB
	TEB00LC	**0.99**	0.94		
	VEB00LC	0.94	**0.99**		
	TEB01LC			**0.99**	0.92
	VEB01LC			0.93	**0.99**

Notes: $\propto$ - Using all the students matched (Schools =482).
β - Using only those students matched in the same school (Schools = 479).
** - All the correlations are significant at the 0.01 level.

The results in Table 10.8 show extremely strong to unity correlations (0.94 to 1.00) within the same subject between the stable (or between the change) Type B and Type C school effects within one data set as well as across the two data sets used. Clearly, most schools show consistent performance across the two types of school effects. That is, the ranking order of the schools based on Type C effects is basically the same as the ranking order of schools based on Type B effects.

In other words, most schools that record more than expected average performance in numeracy (or literacy) based on Type B effects also record more than expected average performance in numeracy (or literacy) based on Type C effects, and vice versa. Similarly, most schools that record more than expected change in performance in numeracy (or literacy) based on Type B effects also record more than expected change in performance in numeracy (or literacy) based on Type C effects, and vice versa. Clearly, within the same data set, inclusion or exclusion of school characteristics (locality and size) and school's mobility rate in the estimation of school effects does not change greatly the ranking order of the primary schools in South Australia based on the stable and change measures of school effectiveness.

Conclusions and discussion

This chapter focuses on Type C school effects. Type C effects are defined as the increment to student achievement attributed to a school after controlling for the effects of student intake and the effects of the average school context and school characteristics. In the longitudinal modelling of the Type C effects, allowances are given for School Size and School Location variables regardless of the statistical significance of their effects because in South Australia these two school characteristics are taken into account by the government when providing funds to the schools. In addition, an allowance is made for the Average Mobility Rate of the school although the effects of this variable do not meet the $p<0.05$ criterion of statistical significance because it is considered that school officials and policy makers might wish to adjust for the effects of mobility.

The main findings in this chapter can be summarized as follows for the two outcome measures of interest in this study.

1. Type C effects are estimated almost as reliably as Type B effects.
2. The stable component of Type C effects is estimated more reliably ($\lambda = 0.37$ to 0.44) than the change component of Type C effects ($\lambda = 0.13$ to 0.16).
3. The stable Type C school effects on numeracy are estimated slightly more reliably than on literacy.
4. There is significant (at $p<0.05$) variability in effectiveness of the primary schools in South Australia based on Type C (both stable and change) indices of school effects but the variance involved is small.
5. Based on the stability ratio criterion (Willms and Raudenbush, 1989), the Type C effects are marginally more stable than the Type B effects.
6. In terms of fixed effects, the results presented in this chapter strongly agree with the results of analyses presented in earlier chapters regarding the student-level and school-level variables that have significant influences on achievement in numeracy and literacy.
7. On the earlier testing occasions, students in large schools are estimated to achieve better in numeracy and in literacy than their counterparts in small schools. However, during the later testing occasions, students in small schools are estimated to achieve better in both subjects than students in large schools.
8. The amounts of variance left unexplained in the Type C effects models at each of the three levels of hierarchy are basically the same as the amounts of variance left unexplained at the corresponding levels of the Type B effects models.

9. Based on the Type C school effects (stable and change) for numeracy and literacy, the ranking of schools obtained using all students who could be matched and the ranking obtained using the students who remained in the same school do not differ markedly ($r \geq 0.94$).
10. The stable Type C effects are to some extent consistent ($r = 0.49$ to 0.55) across the two subjects included in the BSTP. However, the change Type C effects are much less consistent ($r = 0.18$ to 0.23) across the two outcome measures.
11. For numeracy, based on Type C effects, a considerable number of the primary schools that show more than expected average performance are likely to show more than expected increase in performance over time and vice versa ($r = 0.48$ to 0.50). However, for literacy, most of the primary schools that show more than expected average performance are highly likely to show less than expected increase in performance over time and vice versa ($r = -0.71$ to -0.80). These findings for numeracy and literacy are consistent with what is found in the previous chapter for Type A and Type B effects.
12. Within the same subject and the same data set, a vast majority of the primary schools that perform well based on Type B effects (stable and change) also perform well based on Type C effects ($r \geq 0.99$). Thus, it can be concluded that the inclusion or exclusion of school characteristics (namely, school locality and school size) in the estimation of school effects does not change markedly the ranking order of the primary schools in South Australia based on the value added measures of school effectiveness.

At the beginning of this chapter, it was suspected that the amount of variance left at the school-level in Type C effects model would generally be very small, as it has been found to be the case. Thus, despite the potential usefulness of the Type C effects to policy makers and school officials, the variance left unexplained at the school-level is still very small. However, because for the South Australian situation the Type C effects are potentially more useful than the Type B effects, for purposes of ranking schools, it would appear unnecessary to compute the Type B effects. Furthermore, no additional information would be obtained by ranking the schools based on Type B effects because the ranks assigned to the schools based on the Type B effects would not differ greatly from the ranks assigned to the schools based on Type C effects.

Despite what is said above about the dubious value of ranking schools, the procedures described in this chapter could be used to identify unusually effective (i.e. in terms of numeracy and literacy) or unusually ineffective primary schools in South Australia. Those interested could then make on-site visits to schools to identify why the schools are performing well or poorly (Goldstein, 1991; Draper, 1995; Pituch, 1999). For the Type C effects, the on-site visits to schools would focus on monitoring the quality of teaching in the schools, curriculum planning and curriculum implementation in the school, the quality of management in the school and the levels of student motivation and perseverance that are independent of school context.

11

Gender Factor in School Effects

Many researchers have indicated that school effects could be different for different categories of students within a school (for example, Nuttall et al., 1989; Willms and Chen, 1989; Goldstein et al., 1992; Sammons, Nuttall and Cuttance, 1993; Young and Fraser, 1993; Raudenbush and Willms, 1995; Pituch, 1999). Consequently, this chapter addresses the following questions regarding the consistency of school effects for the primary schools in South Australia:

> Are schools that are relatively effective in numeracy for boys also relatively effective for girls?
>
> Are schools that are relatively effective in literacy for girls also relatively effective for boys?

In order to answer the above questions, the longitudinal structure introduced in Chapter 9 is employed in this chapter to estimate indices of individual school effectiveness for boys and girls for the two outcome measures of interest in this study. Hence, for each outcome measure, two indices are computed for each primary school: one for boys and the other for girls. The estimations are carried out using the two data sets: all the students who were matched (Schools = 482); and the students who remained in the same school between Grades 3 and 5 (Schools = 479).

It has been shown in the preceding chapters that the residual variance at the school-level is small and, as a result, it has been argued that it is difficult to judge the relative performance of the primary schools in South Australia based on value added measures. In due course, it has been suggested that some measure of absolute performance needs to be sought. In this regard, an approach to measuring the performance of the school that takes into consideration the time that a student takes to learn certain numeracy (or literacy) skills, is explored in this chapter.

The outline of this chapter is as follows. The first three sections describe the general longitudinal model for the estimation of the school effects for boys and girls, the estimation of these indices of school effectiveness and the results of the HLM analyses

respectively. The fourth to the seventh sections focus on comparing indices of individual school effects for boys and girls.

Specification of the model

In this study, two approaches are considered for the specification of Type A effects for the different categories of students within the school. These two approaches are referred to here as simply (a) 'varying effect' approach (b) 'split-school' approach.

Varying effect approach

Under the varying effect approach, the model for the estimation of Type A effects for each gender is exactly the same as the longitudinal model for the estimation of Type A effects described in Chapter 9. This Type A effects model is identical to the model presented below, except that this time the variable SEX (Sex of the Student) is separated from the other relevant independent variables that describe student's background characteristics, $\boldsymbol{X}_{hitj}$.

Level-1 model

$$\mathbf{Y}_{itj} = \boldsymbol{\pi}_{0tj} + \boldsymbol{\pi}_{1tj}\boldsymbol{SEX}_{itj} + \boldsymbol{\pi}_{htj}\boldsymbol{X}_{hitj} + \mathbf{e}_{itj} \qquad \text{Equation 11.1}$$

Level-2 model

$$\boldsymbol{\pi}_{0tj} = \boldsymbol{\beta}_{00j} + \boldsymbol{\beta}_{01j}\mathbf{OCC}_{tj} + \mathbf{r}_{0tj} \qquad \text{Equation 11.2}$$

$$\boldsymbol{\pi}_{1tj} = \boldsymbol{\beta}_{10j} \qquad \text{Equation 11.3}$$

$$\boldsymbol{\pi}_{htj} = \boldsymbol{\beta}_{h0j} \qquad \text{Equation 11.4}$$

Level-3 model

$$\boldsymbol{\beta}_{00j} = \boldsymbol{\gamma}_{000} + \mathbf{u}_{00j} \qquad \text{Equation 11.5}$$

$$\boldsymbol{\beta}_{01j} = \boldsymbol{\gamma}_{010} + \mathbf{u}_{01j} \qquad \text{Equation 11.6}$$

$$\boldsymbol{\beta}_{10j} = \boldsymbol{\gamma}_{100} + \mathbf{u}_{10j} \qquad \text{Equation 11.7}$$

$$\boldsymbol{\beta}_{h0j} = \boldsymbol{\gamma}_{h00} \qquad \text{Equation 11.8}$$

where:

$\boldsymbol{\pi}_{1tj}$ is the regression coefficient associated with SEX.

All the other components in Equations 11.1 to 11.8 carry the same meaning as described in Chapter 9 for the longitudinal model for the estimation of overall Type A effects. Equations 11.1 to 11.8 can be combined into a single equation to yield the following model, which describes the linear relationship of the components involved.

$$\begin{aligned}\mathbf{Y}_{itj} = {} & [\boldsymbol{\gamma}_{000}] && \text{(grand mean)}\\ & + [\boldsymbol{\gamma}_{010}\mathbf{OCC}_{tj}] && \text{(main effect of occasion)}\\ & + [\boldsymbol{\gamma}_{100}\boldsymbol{SEX}_{it} + \mathbf{u}_{10j}\boldsymbol{SEX}_{itj} + \boldsymbol{\pi}_{htj}\boldsymbol{X}_{hitj}] && \text{(control for student intake)}\\ & + [\mathbf{u}_{00j}] && \text{(stable component of school effect)}\\ & + [\mathbf{u}_{01j}\mathbf{OCC}_{tj} + \mathbf{r}_{0tj}] && \text{(change component of school effect)}\\ & + [\mathbf{e}_{itj}] && \text{(student-level random error)}\end{aligned}$$

Equation 11.9

In the above model (Equation 11.9), the term ($\mathbf{u}_{00j}$) for the stable component and terms ($\mathbf{u}_{01j}\mathbf{OCC}_{tj}$, and $\mathbf{r}_{0tj}$) for the change component are the same as they were in the Type A effects model described in Chapter 9. However, the 'control for student intake' component now includes two more terms: (a) $\boldsymbol{\gamma}_{100}\boldsymbol{SEX}_{it}$, which represent the main effect for gender, and (b) $\mathbf{u}_{10j}\boldsymbol{SEX}_{itj}$, which represent the 'school-by-gender' interaction effect. Of importance here is the school-by-gender interaction effect ($\mathbf{u}_{10j}\boldsymbol{SEX}_{itj}$) because it is the unique influence of school *j* on the achievement of student *i* of a specified gender. Thus, in order to obtain the stable Type A effect for school *j* for a particular gender of students over the study period, the overall stable effect ($\mathbf{u}_{00j}$) and school-by-gender interaction effect ($\mathbf{u}_{10j}\boldsymbol{SEX}_{itj}$) are added (Raudenbush and Willms, 1995; Pituch, 1999). That is:

$$\mathbf{A}_j = \mathbf{u}_{00j} + \mathbf{u}_{10j}\boldsymbol{SEX}_{itj} \qquad \text{Equation 11.10}$$

where:

$\mathbf{A}_j$ is the Type A effect of school *j*.

Because Equation 11.10 includes the variable SEX, it follows that for each school two indices of stable school effects can be computed: one for boys and the other for girls. If in the HLM analyses, each student background variable included in X_{hitj} is grand-mean centred, but the variable SEX (coded; boy = 0, girl = 1) is uncentred, then by substituting in the above equation, A_j for boys is simply:

$$\mathbf{A}_{j(\text{boys})} = \mathbf{u}_{00j} + \mathbf{u}_{10j}^{*}\,[0] = \mathbf{u}_{00j}$$

and, $\mathbf{A}_j$ for girls is:

$$\mathbf{A}_{j(\text{girls})} = \mathbf{u}_{00j} + \mathbf{u}_{10j}^{*}\,[1] = \mathbf{u}_{00j} + \mathbf{u}_{10j}$$

It should be noted, however, that using the above approach, the change effect remains the same for all categories of students in the school, and therefore, it is not possible to examine separately the change effects for the different categories of students within a school.

Split-school approach

Under the split-school approach, the general longitudinal model for estimation of Type A school effects for each gender is exactly the same as the general longitudinal models for estimation of the overall Type A school effects described in the Chapter 9, that is:

$$\begin{aligned}
\mathbf{Y}_{itj} &= [\boldsymbol{\gamma}_{000}] && \text{(grand mean)}\\
&+ [\boldsymbol{\gamma}_{010}\,\mathbf{OCC}_{tj}] && \text{(main effect of occasion)}\\
&+ [\boldsymbol{\pi}_{htj}\boldsymbol{X}_{hitj}] && \text{(control for student intake)}\\
&+ [\mathbf{u}_{00j}] && \text{(stable component of school effect)}\\
&+ [\mathbf{u}_{01j}\mathbf{OCC}_{tj} + \mathbf{r}_{0tj}] && \text{(change component of school effect)}\\
&+ [\mathbf{e}_{itj}] && \text{(student-level random error)}
\end{aligned}$$

Equation 11.11

All the components in Equation 11.11 carry the same meaning as described in Chapter 9 for the longitudinal model for the estimation of overall Type A effects. It is important to remember that $\boldsymbol{\pi}_{htj}\boldsymbol{X}_{hitj}$ represents the control for several relevant independent variables that describe student's background characteristics. More important, it should be borne in mind that the variable SEX is included in the $\boldsymbol{X}_{hitj}$

term and therefore, the effects associated with the student's sex are catered for in the above model.

In order to estimate the school effects for each gender, each school is treated as two separate schools under the split-school approach. That is, two different codes are used to identify each school: one code for boys and the other for girls. Consequently, in the HLM analyses, the numbers of Level-1 units remain the same as they were in the corresponding models described in Chapter 9, that is, 37,832 (when using the transience data set) and 32,741 (when using the non-transience data set). However, because each school is represented twice at Level-2 and twice at Level-3, the numbers of Levels 2 and 3 units in the HLM analyses involving the above approach are twice as many as they were in the corresponding HLM analyses described in Chapter 9. Thus, under the split-school approach it is possible to estimate two pairs of stable and change effects simultaneously: one pair for boys and the other for girls.

The main advantage of the split-school approach over the varying effect approach is that, based on the former approach, it is possible to examine the change effects for the different categories of students within the school. For example, using the split-school approach, it can be examined whether schools that record more than expected change in performance over time for boys also record more than expected change in performance over time for girls.

Based on the split-school approach, there are fewer Level-1 units for the estimation of the parameters at the higher levels compared to the number of Level-1 units involved in the estimation of the parameters at the higher levels based on the varying effect approach. The consequence is that the errors associated with the estimation of the parameters are bound to be larger in the split-school approach than in the varying effect approach. However, if splitting of the schools still leaves an adequate number of students per school, then it would appear to be appropriate to use the split-school approach to estimate the school effects. Through extensive experience, Raudenbush and Willms (1995) argue that sample sizes as low as 25 students per school tend to provide reliable results and under such circumstances the estimated errors can be ignored.

Another problem arises because the split-school approach assumes that gender composition, that is, the ratio of boys to girls in the class has little or no influence on student achievement. However, if gender composition is almost the same across all the schools involved, then the influence of this factor on student achievement could be assumed to be generally the same across all the schools, and therefore, the factor could be ignored.

For both data sets involved in this study, the average is 40 boys per school and 39 girls per school. In addition, the correlation between the number of boys and the number of girls in either of the two data sets is extremely strong (≥ 0.95), indicating that most schools that have many boys also have many girls, and that most schools that have few boys also have few girls. It should be noted that there are neither boys-only schools nor girls-only primary schools involved in this study.

Although the above preliminary data analysis reveals that there is a sufficient number of students per school and fairly similar gender composition across the schools involved, it should be noted, however, that the design for this study is such that there are four data points for the schools. Therefore, splitting of the schools might leave the sample size per occasion less than the 25 students per school recommended by Raudenbush and Willms (1995). The consequence is that it would be difficult to detect differences between schools in their change school effects but it should still be

possible to detect differences between schools in their stable school effects because total numbers of students per school are sufficient.

For the purposes of investigation, both the split-school approach and the varying effect approach are used in this study. Nevertheless, it must be emphasized that, because of the small sample size of students per school on each testing occasion, the results obtained using the split-school approach should be interpreted with some caution. However, the differences on the estimation of the parameters of the models are of sufficient interest to warrant examination of the two approaches.

Estimation of Type A effects for each gender

It has been noted above that to estimate Type A effects for boys and girls, the split-school approach and the varying effect approach are employed. The estimation is carried out first using the transience data set and then using the non-transience data set.

For the varying effect approach, the final models used to estimate the Type A effects in this chapter are basically identical to the final models used to estimate Type A effects in Chapter 9.

At Level-1 and Level-2, the final models employed to estimate Type A effects using the split-school approach are identical to the final models used to estimate the Type A effects in Chapter 9. At Level-3, the effects associated with the student's sex can only be specified as fixed within the split-school approach because each Level-3 unit is treated as either a boy-school or a girl-school. This procedure is contrary to the analyses described in Chapter 9 where in all models the effects associated with the student's sex were found to vary significantly across the schools and were, therefore, specified as random.

Results

The results of HLM analyses described above for the estimation of Type A effects for boys and girls provide estimates of reliability, fixed effects, variance components and the deviance statistics.

The first panel of Table 11.1 displays the school-level reliability estimates of the stable and the change components from the simplest longitudinal models using the varying effect approach, while the second panel displays the corresponding information using the split-school approach. As it would be expected, the splitting of schools into boys and girls schools leads to a general decline in the reliability estimates of the stable and the change components. The reduction in the reliability estimates is because there are now fewer numbers of students per school in the split-school analyses compared to the numbers of students per school in the varying effect analyses. Nevertheless, the results displayed in Table 11.1 show that the reliability estimates of both the stable and change components for the split-school approach are all above 0.05 level (Bryk and Raudenbush, 1992; Pituch, 1999).

For the varying effect approach, the results of the above HLM analyses are basically identical to the results that are obtained in Chapter 9 from the corresponding Type A effects models and, therefore, no additional discussion of these results is necessary here (see Chapter 9).

For the split-school approach, apart from the general decline in the reliability estimates of the stable and change components, the other results of the above HLM analyses follow closely the results obtained in Chapter 9 from the corresponding Type

A effects models. Again, no additional discussion of the results of these HLM analyses is necessary here.

The following sections concentrate on the similarities and the differences between the indices of individual school effects for boys and girls computed using the two approaches described above. The sections focus on (a) the correlations between the indices of individual school effects, (b) the descriptive statistics of the school effects, and (c) the patterns of schools effects across gender.

Table 11.1 School-level reliability estimates from simplest longitudinal models using varying effect and split-school approaches

			Numeracy		Literacy	
Approach			Tran$^{\propto}$	Non-Tran$^{\beta}$	Tran$^{\propto}$	Non-Tran$^{\beta}$
	Effect	***Random level-2 coefficient***				
Varying effect	Stable	INTRCPT1/ INTRCPT2,	0.859	0.836	0.844	0.825
	Change	INTRCPT1/ OCC,	0.092	0.053	0.106	0.116
Split-school	Stable	INTRCPT1/ INTRCPT2,	0.769	0.738	0.776	0.752
	Change	INTRCPT1/ OCC,	0.058	0.052	0.066	0.053

Notes: $\propto$ - Using all the students matched (Schools = 482).
β - Using only those students matched in the same school (Schools = 479).

Correlation between varying effect and split-school stable school effects

For presentation purposes, the same system used in the naming of school effects in the preceding chapters is used here. Hence, the codes that start with a 'T' are for school effects obtained using the transience data set while the codes that start with a 'V' are those for school effects obtained using the non-transience data set. The prefixes 'EB00' and 'EB01' represent the empirical Bayes estimates for the stable and the change components of school effects respectively. However, in order to differentiate between the school effects for boys and girls, 'M' (male) or 'F' (female) is included at the end of a code. Thus, an 'NAM' ending indicates that the code represent Type A effects for numeracy for boys while an 'LAF' ending indicates that the code represent Type A effects for literacy for girls. For example, in Table 11.2, the codes 'TEB00NAM' and 'TEB00LAM' represent the stable Type A school effects for numeracy and literacy for boys obtained using the transience data set respectively.

Table 11.2 displays the correlations between the stable school effects computed using the varying effect approach and those computed using the split-school approach. Under the varying effect approach, the change effect remains the same for all categories of students in the school, and therefore, it is not possible to examine correlations between the change effects obtained using the two approaches.

The results in Table 11.2 show extremely strong correlations (0.89 to 0.94) between the stable school effects obtained using these two approaches, regardless of the data set used. Thus, within the same gender of students, the ranking order of the schools based on stable Type A effects obtained using the varying effect approach does not differ markedly from the ranking order of schools based on stable Type A effects obtained using the split-school approach.

Table 11.2 Correlation between varying effect and split-school stable school effects

	Data	Variable	Correlation
Numeracy	Tran$^{\alpha}$	TEB00NAM	0.94
		TEB00NAF	0.90
	Non-Tran$^{\beta}$	VEB00NAM	0.93
		VEB00NAF	0.90
Literacy	Tran$^{\alpha}$	TEB00LAM	0.90
		TEB00LAF	0.89
	Non-Tran$^{\beta}$	VEB00LAM	0.90
		VEB00LAF	0.90

Notes: $\propto$ - Using all the students matched (Schools = 482).
β - Using only those students matched in the same school (Schools = 479).

Correlations between school effects for boys and girls

The top panel of Table 11.3 displays the correlations between stable Type A effects for boys and the stable Type A effects for girls obtained using the varying effect approach. The bottom panel of Table 11.3 displays the correlations between Type A (both stable and change) effects for boys and the Type A effects for girls (both stable and change) obtained using the split-school approach. Under the varying effect approach the change effect remains the same across all the categories of the students within the school and therefore it is not possible to examine the correlations between the change effects under this approach.

Each panel of Table 11.3 displays the correlations obtained using the transience data set as well as those obtained using the non-transience data set for the two outcome measures of interest in this study. The numbers given in bold in the table are the correlations between the stable (or between the change) Type A effects for boys and the stable (or change) Type A effects for girls within the same subject and for the same data set obtained using the same approach.

In interpreting the results in Table 11.3, it should be remembered that the varying effect results are estimated more reliably compared to the split-school results. For the varying effect approach, the results in Table 11.3 show extremely strong correlations ($\geq$0.95) between the stable school effects for boys and girls within one data set as well as across the two data sets used. The extremely strong correlations indicate that, using the varying effect approach, almost all the schools that record more than expected average performance in numeracy (or literacy) for boys also record more than expected average performance for girls in numeracy (or literacy), and vice versa. More important, within the same data set and using the varying effect approach, the ranking order of the schools based on stable Type A effects for boys is basically the same as the ranking order of schools based on stable Type A effects for girls ($r \geq$ 0.98) for both numeracy and literacy.

For the split-school approach, the results in Table 11.3 show very strong correlations (0.71 to 0.77) between the stable school effects and large correlations (0.57 to 0.66) between change school effects for boys and girls within one data set as well as across

the two data sets used. For the stable effects, the very strong correlations indicate that using the split-school approach, a vast majority of schools that record more than expected average performance in numeracy (or literacy) for boys also record more than expected average performance for girls in numeracy (or literacy), and vice versa. For the change effects, the large correlations indicate that a considerable number of schools that record more than expected change in performance in numeracy (or literacy) over time for boys also recorded more than expected change in performance in numeracy (or literacy) over time for girls, and vice versa.

Table 11.3 Correlations between Type A school effects for boys and girls

			For Girls			
			Stable		**Change**	
			Tran$^{\propto}$	Non-Tran$^{\beta}$	Tran$^{\propto}$	Non-Tran$^{\beta}$
For Boys						
	Varying effect	Numeracy**				
			TEB00NAF	VEB00NAF		
		TEB00NAM	**0.98**	0.95		
		VEB00NAM	0.97	**0.98**		
		Literacy**				
			TEB00LAF	VEB00LAF		
		TEB00LAM	**0.98**	0.96		
		VEB00LAM	0.96	**0.99**		
	Split-school	Numeracy**				
			TEB00NAF	VEB00NAF	TEB01NAF	VEB01NAF
		TEB00NAM	**0.77**	0.76		
		VEB00NAM	0.76	**0.75**		
		TEB01NAM			**0.66**	0.59
		VEB01NAM			0.62	**0.60**
		Literacy**				
			TEB00LAF	VEB00LAF	TEB01LAF	VEB01LAF
		TEB00LAM	**0.74**	0.73		
		VEB00LAM	0.71	**0.71**		
		TEB01LAM			**0.61**	0.60
		VEB01LAM			0.57	**0.57**

Notes: $\propto$ - Using all the students matched (Schools = 482).
β - Using only those students matched in the same school (Schools = 479).
** - All the correlations are significant at the 0.01 level.

From the results in Table 11.3, it is clear that the correlations between the stable school effects obtained using the split-school approach are considerably smaller when compared to the correlations between the stable school effects obtained using the varying effect approach. Because the varying effect results are estimated more reliably than the split-school results, it is likely that the varying effect results provide a better

picture of the relationships involved compared to the split-school results. Furthermore, it would generally be expected that at the lower grade levels most primary schools that are effective for boys would also be effective for girls and, therefore, the correlations between the stable school effects for boys and girls would be as strong as is observed using the varying effect approach. Following the same argument, it is likely that the correlation between change effects for boys and girls could be much stronger than the correlations observed using the split-school approach (results in Table 11.3).

Descriptive statistics

The top panel of Table 11.4 displays the descriptive statistics for stable school effects by gender obtained using the varying effect approach, while the second panel displays the descriptive statistics for both stable and change school effects by gender obtained using the split-school approach. Each panel of Table 11.4 displays the descriptive statistics for the school effects obtained using the transience data set (Schools = 482) as well as using the non-transience data set (Schools = 479).

When interpreting the results displayed in Table 11.4, it is important to consider the following three issues. First, in general, zero is the average of the school effects and schools with values with positive signs are considered to be relatively effective when compared to the average while schools with values with negative signs are considered to be relatively ineffective when compared to the average. That is, schools with positive values are likely to contribute more to the increase in student achievement, while schools with negative values tend to contribute less to the increase in student achievement.

Second, the sizes of school effects depend on the nature of the distribution of the outcome variable. Therefore, it is misleading to compare the values of school effects across outcome measures without adjusting to take account of the differences in the distribution of the outcome variables. Willms (1992) and Harker and Nash (1996) argue that expressing the school effect as a fraction of the standard deviation of the outcome measure can do this adjustment. The resulting adjustment produces what Willms (1992; p.43) calls an "effect size" for each school, which can now be compared across outcome measures. For the current study, the standard deviations for numeracy and literacy when using the transience data set are 1.40 and 1.50 respectively, and when using the non-transience data set are 1.44 and 1.54 respectively. Consequently, to compare across the outcome measures the minimum, maximum and range values displayed in Table 11.4 have first to be divided by the corresponding standard deviations. For the range, the resulting effect sizes are shown in Table 11.4.

Third, the current HLM software does not provide standard errors of the school-effects, and therefore, the precision of the school effects is unknown (Pituch, 1999). More important, without knowledge of the standard errors associated with the school effects the statistical significances of the differences between the school effects are likely to be flawed. Consequently, in order to judge the importance of the difference between the school effects in this study, faith is placed on the estimated average growth in numeracy and literacy achievement between the Grades 3 and 5. The average growth in numeracy (or literacy) achievement between Grades 3 and 5 has been estimated as about 0.50 logits per year (Hungi, 1997; see also Chapters 6 and 7 of the current study). As a result, a Type A school effect of 0.125 (about 0.13) logits would indicate that after controlling for the student background characteristics, the school's contribution to an increase in student achievement is about one school-term's work and, therefore, is substantial.

Table 11.4 Descriptive statistics of stable and change school effects by gender

					Mean	Std. Dev.	Min.	Max.	Range	E. Size Range
Varying effect	Stable	Numeracy	Tran$^{\propto}$	TEB00NAM	0.00	0.18	-0.40	0.49	0.89	0.64
				TEB00NAF	0.00	0.17	-0.38	0.44	0.81	0.58
			Non-Tran$^{\beta}$	VEB00NAM	0.00	0.17	-0.42	0.44	0.86	0.60
				VEB00NAF	0.00	0.16	-0.43	0.43	0.86	0.59
		Literacy	Tran$^{\propto}$	TEB00LAM	0.00	0.13	-0.45	0.33	0.79	0.53
				TEB00LAF	0.00	0.12	-0.42	0.29	0.71	0.47
			Non-Tran$^{\beta}$	VEB00LAM	0.00	0.11	-0.35	0.29	0.64	0.41
				VEB00LAF	0.00	0.10	-0.34	0.28	0.62	0.40
Split-school	Stable	Numeracy	Tran$^{\propto}$	TEB00NAM	0.00	0.16	-0.43	0.45	0.88	0.63
				TEB00NAF	0.00	0.14	-0.40	0.48	0.88	0.63
			Non-Tran$^{\beta}$	VEB00NAM	0.00	0.14	-0.44	0.45	0.89	0.62
				VEB00NAF	0.00	0.14	-0.41	0.46	0.87	0.60
		Literacy		TEB00LAM	0.00	0.11	-0.41	0.33	0.74	0.49
				TEB00LAF	0.00	0.10	-0.36	0.31	0.67	0.45
				VEB00LAM	0.00	0.09	-0.39	0.26	0.65	0.42
				VEB00LAF	0.00	0.09	-0.33	0.28	0.61	0.40
	Change	Numeracy	Tran$^{\propto}$	TEB01NAM	0.00	0.01	-0.04	0.04	0.08	0.06
				TEB01NAF	0.00	0.01	-0.03	0.05	0.08	0.06
			Non-Tran$^{\beta}$	VEB01NAM	0.00	0.01	-0.04	0.04	0.08	0.06
				VEB01NAF	0.00	0.01	-0.04	0.05	0.09	0.06
		Literacy	Tran$^{\propto}$	TEB01LAM	0.00	0.02	-0.05	0.08	0.13	0.09
				TEB01LAF	0.00	0.02	-0.06	0.06	0.12	0.08
			Non-Tran$^{\beta}$	VEB01LAM	0.00	0.02	-0.05	0.08	0.13	0.08
				VEB01LAF	0.00	0.02	-0.06	0.07	0.13	0.08

Notes: $\propto$ - Using all the students matched (Schools = 482).
β - Using only those students matched in the same school (Schools = 479).
Min. - Minimum.
Max. - Maximum
E.Size - Effect Size

For the stable Type A school effects, a difference of around ≥0.13 (range, in Table 11.4) between effects of the most effective school (maximum, in Table 11.4) and the least effective school (minimum, in Table 11.4) should be considered substantial. Clearly, within the same gender, for both subject areas, the information displayed in Table 11.4 indicates that there are substantial differences in the stable Type A school effects, regardless of the approach employed. For example, for girls while using the non-transience data set, the differences between the most effective school and the least effective school for numeracy (VEB00NAF) are estimated to be 0.86 and 0.87 logits when using the varying effect approach and the split-school approach respectively.

For this example, the differences are about seven terms of schoolwork. These differences (that is, 0.86 and 0.87 logits) in the contributions made to the increase in student achievement between the most effective and the least effective schools should also be considered substantial in relation to the grand mean for numeracy, which is estimated at around 1.42 logits using the varying effect approach.

Thus, using the 0.13 logits as an indicator of one school-term of work in numeracy and literacy, it is easy to grasp the importance of the differences between the most effective school and the least effective schools for the stable effects. For the change school effect, however, the judgement of the importance of the range is not so straightforward.

It should be considered that the change effect of a school reflects the changes that have occurred in the intake-adjusted performance level of the school over the study period with each school serving as its own control (Willms and Raudenbush, 1989; p. 214). Thus, compared to what is expected of the school, a positive value of the change effect indicates that the school's intake-adjusted performance level increases while a negative value indicates that the school's intake-adjusted performance level decreases. Consequently, any school (effective or ineffective) based on the stable effects criterion could have either a positive, a zero or a negative value of the change effect depending on whether the school's intake-adjusted performance level increases, remains the same, or decreases over the study period respectively.

Because the maximum and minimum values of change effects in Table 11.4 are not necessarily from schools with equal values of stable effects, it is difficult to make a judgement regarding the impact of the range to the overall difference in the contributions of the schools to increase in student achievement. Furthermore, even for schools with equal values of stable school effects, a typical so-called 'fan effect' pattern would be observed when a school with a positive value of change effects is compared to a school with a negative value of change effect. That is, the gap between the school with a positive value of the change effect and a school with a negative value of the change effect would increase over the testing occasion. Consequently, it is not easy to establish how large the range of the change school effects should be for it to be considered as substantial.

Finally, the results in Table 11.4 indicate that within the same outcome measure, there is some degree of consistency between the minimum, maximum and, consequently, the range values across the two approaches, as well as across the two data sets and gender. In addition, the results in Table 11.4 (effect size range) indicate that regardless of the gender and regardless of the approach used, the differences between the most and the least effective schools based on the stable effects criteria for numeracy are slightly greater when compared to the corresponding differences for literacy. For example, when using the split-school approach, the effect size range values for the stable school effects for numeracy are (0.60 to 0.63) slightly greater than the corresponding values for literacy (0.40 to 0.49). Similarly, the standard deviation for school effects on numeracy are slightly larger than the standard deviation for school effects on literacy, which seems to suggests that there is more variation between the effectiveness of the primary school in numeracy than in literacy. However, within the same subject area, for both stable and change effects, the effect size range observed for boys follows closely the range observed for girls.

Patterns of school effects

This section examines the patterns of school effects for boys and girls within individual schools. However, for reasons of parsimony, this sub-section focuses only

on the examination of the patterns of stable school effects obtained using the varying effect approach and the transience data set (Schools = 482) for schools that:

(a) have the largest differences in school effects between boys and girls; and

(b) are on the extreme ends of the ranks for boys and girls.

Schools with the largest gender differences in effects

Panels 1 and 2 of Table 11.5 give summaries of the number of schools that record differences in school effects between boys and girls that can be considered substantial (greater than 0.13 logits) for numeracy and literacy respectively, obtained using the varying effect approach. Each panel of Table 11.5 also gives the number of schools that record differences in school effects that are greater than 0.063 logits (about a half of school-term's work).

For example, when considering the transience data set, panel 1 shows that the number of schools that record differences in school effects in numeracy that are greater than 0.063 in favour of boys are 14 (2.9 per cent), and in favour of girls, are 11 (2.3 per cent), giving a total of 25 (5.2 per cent). Of these 14 schools that record differences greater than 0.063 in favour of boys, only one school records a difference that is greater than 0.125 logits, while five out of the 11 schools that record differences greater than 0.063 in favour of girls record differences that are greater than 0.125 logits.

Table 11.5 Schools with substantial differences in effects in favour of one gender group

	Differences in	**Boys**		**Girls**		**Total**	
	School Effect	N	%	N	%	N	%
Numeracy							
Tran$^{\propto}$	>0.063	14	2.9	11	2.3	25	5.2
	>0.125	1	0.2	5	1.0	6	1.2
Non-Tran$^{\beta}$	>0.063	12	2.5	9	1.9	21	4.4
	>0.125	1	0.2	4	0.8	5	1.0
Literacy							
Tran$^{\propto}$	>0.063	4	0.8	6	1.2	10	2.1
	>0.125	0	0.0	3	0.6	3	0.6
Non-Tran$^{\beta}$	>0.063	0	0.0	4	0.8	4	0.8
	>0.125	0	0.0	0	0.0	0	0.0

Notes: $\propto$ - Using all the students matched (Schools = 482).
β - Using only those students matched in the same school (Schools = 479).

In summary, the results in Table 11.5 show that schools that record differences in school effects that are greater than half a term of schoolwork in favour of either boys or girls, are 4.4 to 5.2 per cent for numeracy, and for literacy, they are 0.8 to 2.1 per cent. The results in Table 11.5 also show that, in general, very few schools (about one per cent or less) record difference in school effects greater than 0.125 logits in favour of either boys or girls.

Figures 11.1 and 11.2 show the histogram plots of the school effects for the schools that record differences in school effects between boys and girls that could be considered substantial for numeracy and literacy respectively. For purposes of

illustration only, the top ten schools using the varying effect approach are included in each of the histogram plots.

In order to maintain the confidentiality of the findings, the naming system used to identify the schools here is different from the one used by DETE. Under the naming system used here, each school is allocated a code that starts with 's'.

In the histogram plots, the numbers of boys and girls in each school are given in parenthesis after the school's code. Thus, s033 (62, 54) in Figure 11.1 means that in school s033 a total of 62 boys and 54 girls are involved in the computation of the school effects for this school using this data set (that is, the transience data set). Clearly, based on the criterion of at least 25 students per school (Raudenbush and Willms, 1995), all the schools included in Figures 11.1 and 11.2 have sufficient numbers of boys and girls for reliable estimation of the school-level parameters under the varying effect approach.

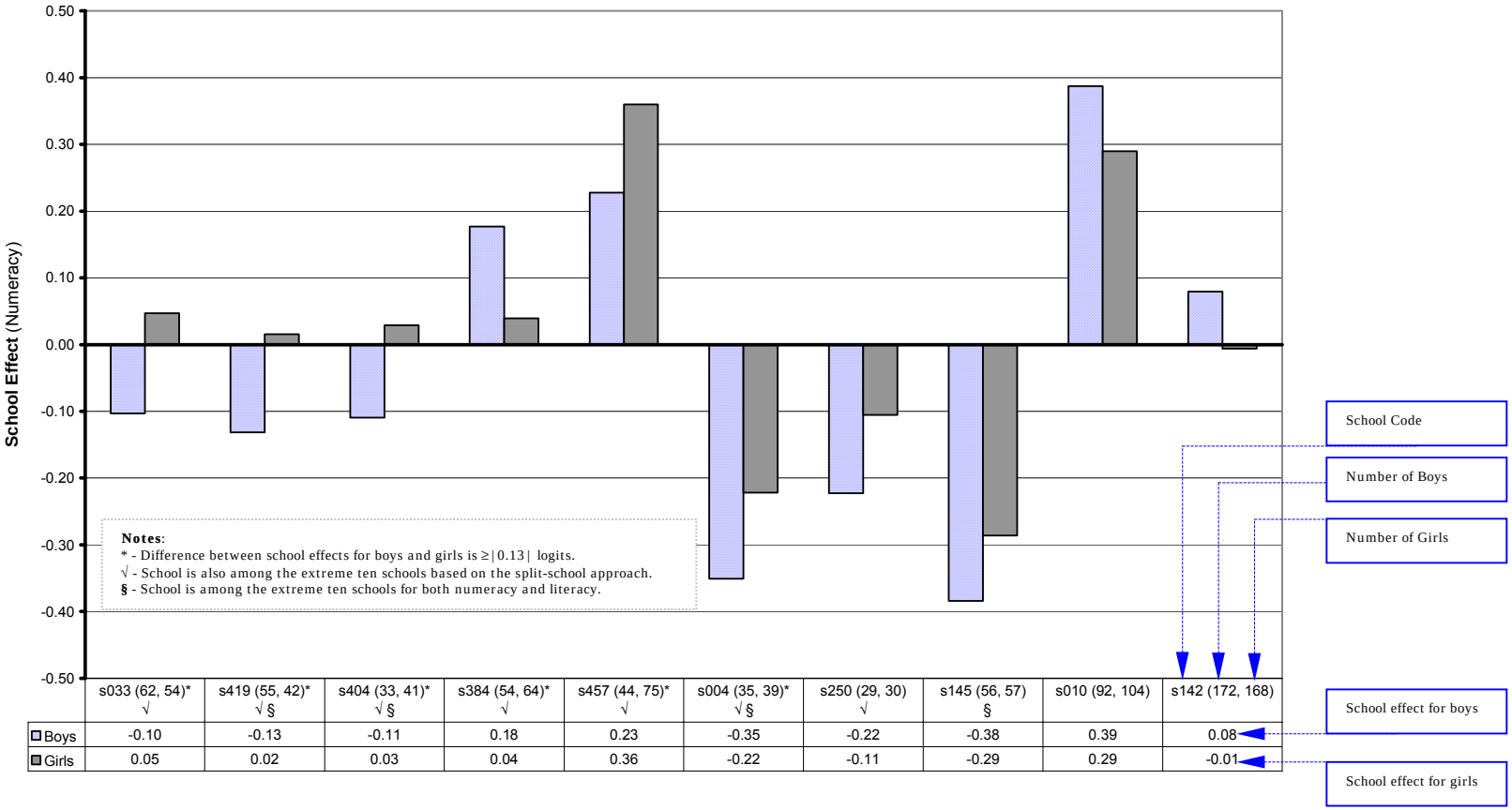
In addition, in the histograms, an asterisk (*) after the school code is used to indicate that the absolute difference between the school effects for boys and girls is equal to or exceeds one school-term's work. A tick symbol (√) appearing below the school code is used to indicate that the school is also among the top ten schools when considering the results from the split-school approach. The symbol § is used to indicate that the school is among the top ten schools for both numeracy and literacy. And finally, the figures in the data tables at the bottom of histograms are the estimates of the school effects for boys and girls for each school included in the plot.

For example, the first two columns in Figure 11.1 show that school s033 is relatively ineffective in numeracy for boys (-0.10) and relatively effective in numeracy for girls (0.05), that overall makes this school more effective in numeracy for girls than for boys by about one school-term's work (0.15).

The last two columns in Figure 11.1 show that the reverse pattern is the case for school s142; that is, school s142 is more effective in numeracy for boys (0.08) than for girls (-0.01) by about three quarters of a school-term's work (0.09). As another example, schools s004 and s457 are overall more effective in numeracy for girls than for boys by about one school-term's work. However, school s004 is relatively ineffective in numeracy for boys (-0.35) as well as for girls (-0.22), while school s457 is relatively effective in numeracy for boys (0.23) as well as for girls (0.36).

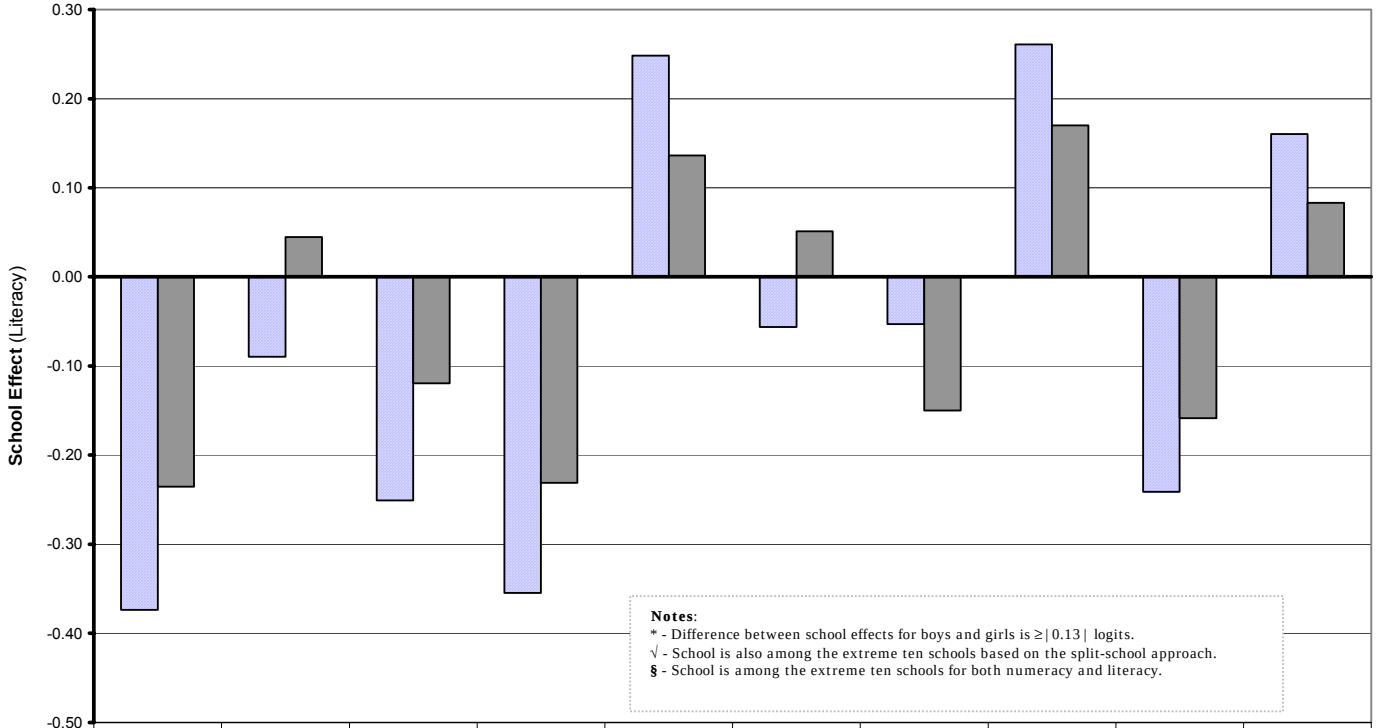
Four schools (namely, s004, s145, s404 and s419) among the top ten schools that recorded the largest differences in school effects between boys and girls for numeracy (Figure 11.1) are among the top ten schools that recorded the largest differences in school effects between boys and girls for literacy (Figure 11.2). Interestingly, one of these four schools (s419) is relatively effective for girls (that is, has positive effects), yet it is relatively (that is, has negative effects) and substantially (that is, difference $\geq|0.13|$) ineffective for boys. Clearly, apart from the other schools illustrated in Figures 11.1 and 11.2, administrators and superintendents would be interested in finding out why this school (s419) appears to be effective in numeracy (and literacy) for one gender yet very ineffective for the other gender.

Finally, for numeracy, the tick symbols (√) in Figure 11.1 show that seven out of the top ten schools that record the largest differences between the school effects for boys and girls when the varying effect approach is used, also record the largest differences when the split-school approach is used. Likewise, for literacy (Figure 11.2), five schools are among the top ten schools that record the largest differences between the school effects regardless of the approach used, which shows some degree of consistency in the results obtained using the two approaches.



	s033 (62, 54)* √	s419 (55, 42)* √ §	s404 (33, 41)* √ §	s384 (54, 64)* √	s457 (44, 75)* √	s004 (35, 39)* √ §	s250 (29, 30) √	s145 (56, 57) §	s010 (92, 104)	s142 (172, 168)
Boys	-0.10	-0.13	-0.11	0.18	0.23	-0.35	-0.22	-0.38	0.39	0.08
Girls	0.05	0.02	0.03	0.04	0.36	-0.22	-0.11	-0.29	0.29	-0.01

Figure 11.1 Schools with largest gender differences in effectiveness in numeracy



	s367 (38, 55)* √	s419 (55, 42)* √ §	s004 (35, 39)* √ §	s145 (56, 57) §	s245 (183, 176)	s133 (58, 58) √	s255 (48, 60) √	s241 (110, 88)	s404 (33, 41) §	s427 (64, 89)
Boys	-0.37	-0.09	-0.25	-0.35	0.25	-0.06	-0.05	0.26	-0.24	0.16
Girls	-0.24	0.04	-0.12	-0.23	0.14	0.05	-0.15	0.17	-0.16	0.08

Figure 11.2 Schools with the largest gender differences in effectiveness in literacy

Most effective and least effective schools in numeracy

Figures 11.3 and 11.4 show the histogram plots of the school effects (obtained using the varying effect approach) for the most effective schools in numeracy for boys and girls respectively. Figures 11.5 and 11.6 show the corresponding plots for the least effective schools. Again, for purposes of illustration, only the top ten schools are included in Figures 11.3 and 11.4, and only the bottom ten schools are included in Figures 11.5 and 11.6. The corresponding histogram plots for literacy can be found in Appendix 14.7.

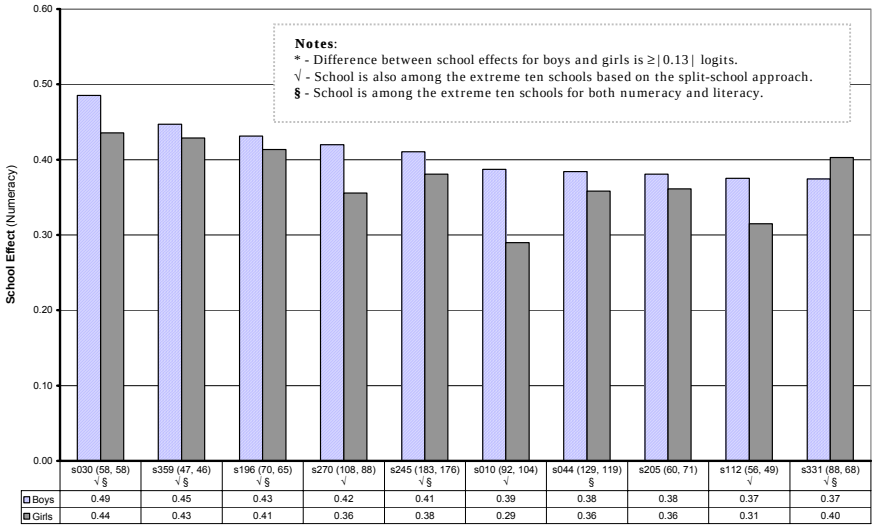
Figure 11.3 illustrates that most schools that are relatively effective in numeracy for boys are also relatively effective in numeracy for girls. Similarly, Figure 11.4 illustrates that most schools that are relatively effective in numeracy for girls are also relatively effective in numeracy for boys. And Figure 11.5 illustrates that most schools that are relatively ineffective in numeracy for boys are also relatively ineffective in numeracy for girls. Likewise, for numeracy, Figure 11.6 illustrates that most schools that are ineffective for girls are also ineffective for boys.

In addition, it should be noted that eight schools (s030, s044, s196, s205, s245, s270, s331 and s359) that are among the top ten effective schools for boys (Figure 11.3) are also among the top ten effective schools for girls in numeracy (Figure 11.4). At the other end, seven schools (s237, s255, s333, s346, s345, s351 and s461) that are among the bottom ten effective schools for boys (Figure 11.5) are also among the bottom ten effective schools for girls in numeracy (Figure 11.6).

Thus, the illustrations in Figures11.3 to 11.6 seem to confirm what was found earlier in the chapter, that is, there are strong correlations ($r \geq 0.95$) between school effects for boys and girls. However, if the aim is to identify schools that have differential effects, then the histograms have an advantage over mere correlation coefficients. The histograms provide information regarding the differences between school effects for boys and girls in individual schools and, therefore, potential 'outlier' schools can be identified for further scrutiny. For example, the $\geq|0.13|$ logits criterion could be used to define the 'outlier' schools and, consequently, schools recording differences between school effects for boys and girls outside this limit could be listed for further examination, regardless of the overall relative effectiveness of the schools.

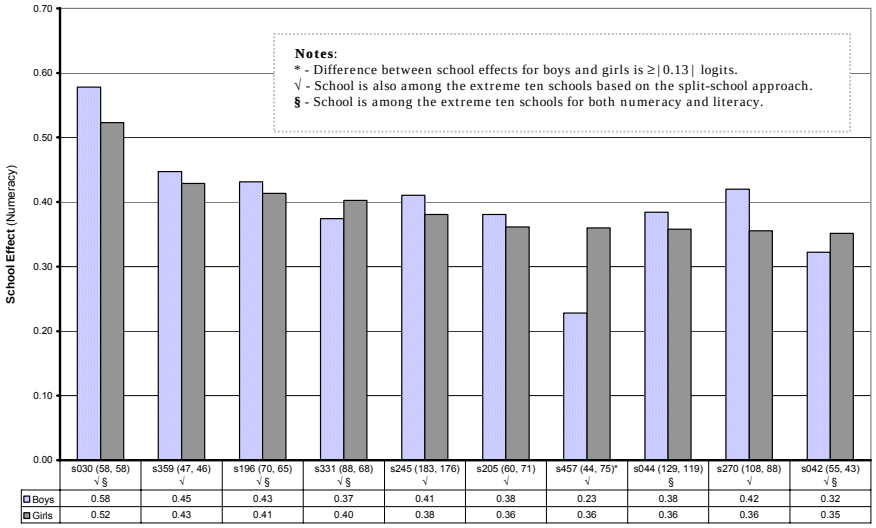
Thus, for the illustration provided in this section, the following seven schools could be listed for further scrutiny based on the $\geq|0.13|$ criterion: s004, s033, s367, s384, s404, s419 and s457. Perhaps, administrators and superintendents would be more interested in school s419 because, for both numeracy and literacy it appears to be relatively effective for girls (0.02 and 0.04 for numeracy and literacy respectively) and substantially ineffective for boys (-0.13 and -0.09 for numeracy and literacy respectively). Another school that should attract interest is school s457 because although it is relatively effective for both boys and girls in numeracy, it nevertheless appears to be substantially more effective for girls than for boys (Figure 11.4). Similarly, school s004 should also attract interest because it appears to be consistently ineffective for both boys and girls, but substantially more ineffective for boys than for girls (Figures 11.1 and 11.2). (See also histogram plots for literacy in Appendix 14.7).

Finally, it should be noted from the tick symbols (√) in Figures 11.3 to 11.6 that most schools that are identified to be among the extreme ten (top or bottom) schools based on the varying effect approach are also identified to be among the extreme ten schools based on the split-school approach. Likewise, from the § symbols it should be noted that a considerable number of schools that are identified to be at the extreme for numeracy are also identified to be at the extreme for literacy.



	s030 (58, 58) √ §	s359 (47, 46) √ §	s196 (70, 65) √ §	s270 (108, 88) √	s245 (183, 176) √ §	s010 (92, 104) √	s044 (129, 119) §	s205 (60, 71)	s112 (56, 49) √	s331 (88, 68) √ §
Boys	0.49	0.45	0.43	0.42	0.41	0.39	0.38	0.38	0.37	0.37
Girls	0.44	0.43	0.41	0.36	0.38	0.29	0.36	0.36	0.31	0.40

Figure 11.3 Top ten effective schools for boys in numeracy



	s030 (58, 58) √ §	s359 (47, 46) √	s196 (70, 65) √ §	s331 (88, 68) √ §	s245 (183, 176) √	s205 (60, 71) √	s457 (44, 75)* √	s044 (129, 119) §	s270 (108, 88) √	s042 (55, 43) √ §
Boys	0.58	0.45	0.43	0.37	0.41	0.38	0.23	0.38	0.42	0.32
Girls	0.52	0.43	0.41	0.40	0.38	0.36	0.36	0.36	0.36	0.35

Figure 11.4 Top ten effective schools for girls in numeracy

Conclusions

This chapter focuses on gender differences in Type A school effects for primary schools in South Australia. The longitudinal model introduced in Chapter 9 is employed in this chapter to estimate indices of individual school effectiveness for boys and girls for numeracy and literacy using two approaches: (a) varying effect approach, and (b) split-school approach.

Under the varying effect approach, the stable school effect for school *j* for a particular gender of students is obtained by adding up the overall stable effect ($\mathbf{u}_{00j}$) and school-by-gender interaction effect ($\mathbf{u}_{10j}\boldsymbol{SEX}_{ij}$). Under the split-school approach, school *j* is

treated as two separate units (one for boys and the other for girls) in order to obtain both the stable and the change indices of individual school effectiveness for each gender.

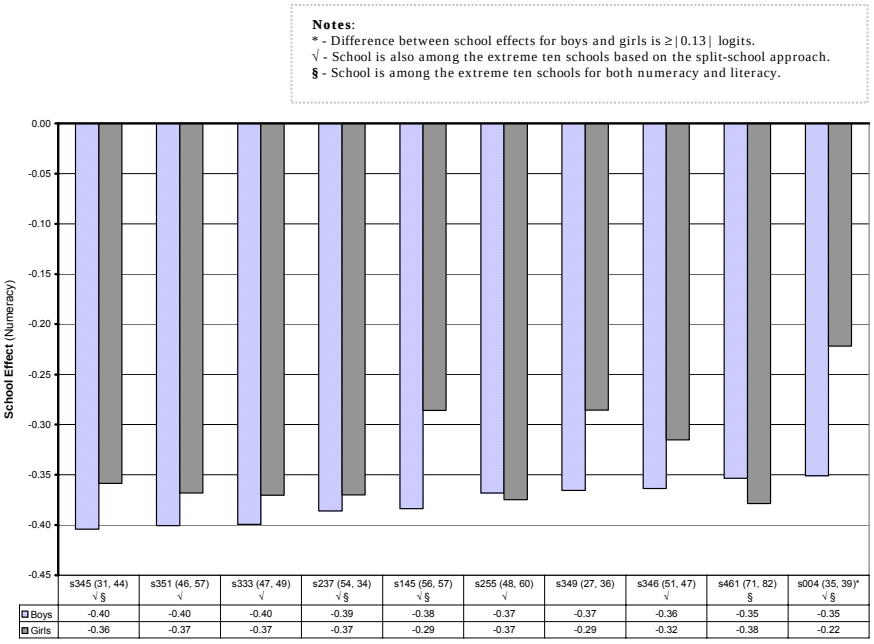


Figure 11.5 Ten least effective schools for boys in numeracy

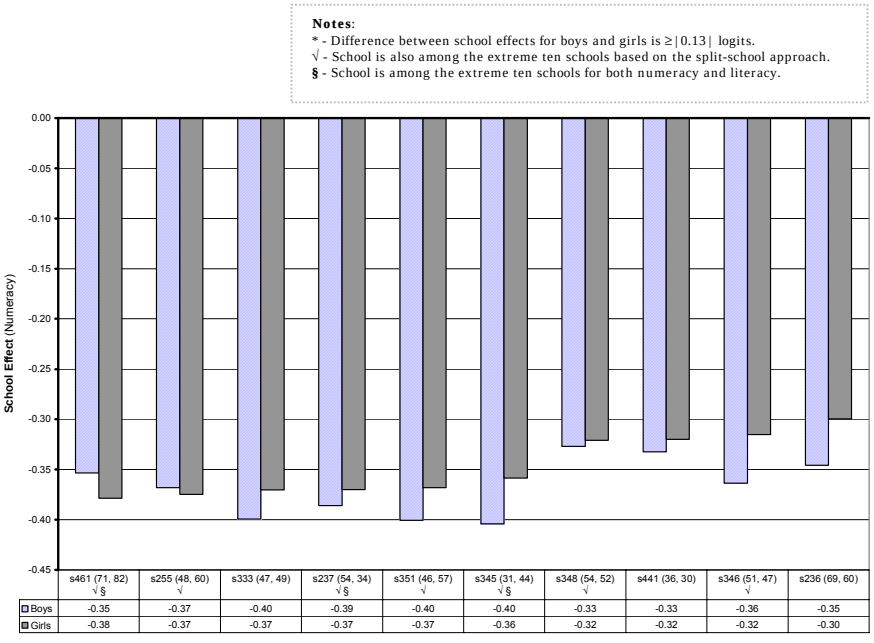


Figure 11.6 Ten least effective schools for girls in numeracy

The main findings in this chapter can be summarized as follows for the two outcome measures of interest in this study.

1. School-level parameters are estimated more reliably under the varying effect approach than they are under the split-school approach.
2. Within the same gender group of students, the rank order of the schools based on stable Type A effects obtained using the varying effect approach does not differ markedly from the rank order of schools based on stable Type A effects obtained using the split-school approach (r = 0.89 to 0.94).
3. Within the same outcome measure, the stable Type A school effects are highly consistent across gender categories (varying effect $r \geq 0.95$; split-school r = 0.71 to 0.77). That is, a vast majority of schools that record more than expected average performance in numeracy (or literacy) for boys also record more than expected average performance for girls in numeracy (or literacy), and vice versa. Furthermore, within the data set and using the varying effect approach, the rank order of the schools based on stable Type A effects for boys is basically the same as the rank order of schools based on stable Type A effects for girls ($r \geq 0.98$) for both numeracy and literacy.
4. Within the same outcome measure, the change Type A school effects are fairly consistent (split-school r = 0.57 to 0.66). That is, a considerable number of schools that records more than expected change in performance in numeracy (or literacy) over time for boys also record more than expected change in performance in numeracy (or literacy) over time for girls, and vice versa.
5. Based on the effect size criterion (Willms, 1992; p. 43) it would appear that:
   (a) within the same subject area and gender, for the stable effects, the difference between the most and the least effective schools under the varying effect approach follow closely the difference between the most and the least effective schools under the split-school approach;
   (b) within the same subject area, for both stable and change effects, the difference between the most and the least effective schools for boys follow closely the difference between the most and the least effective schools for girls; and
   (c) for the stable school effects, there are more variations between the effectiveness of the primary schools in numeracy than in literacy.
6. For the stable school effects, based on a criterion level of 0.13 logits as an indication of the amount of learning done in numeracy and literacy within one school-term:
   (a) within the same gender, there are substantial differences between the most effective schools and the least effective schools in numeracy as well as in literacy;
   (b) across the gender groups, some schools can be differentially and substantially effective (or ineffective) in favour of either boys or girls;
   (c) some schools that are effective for one gender of students can be substantially ineffective for the other gender of students in both numeracy and literacy; and
   (d) a vast majority of schools that are identified as so-called ‘outliers’ under the varying effect approach are also identified as outliers under the split-school approach.

As a note to clarify 6(b) above, around 4.4 to 5.2 per cent and around 0.8 to 2.1 per cent of the schools included in this study record differences in school effects that are greater than half a term of schoolwork in favour of either boys or girls in numeracy and literacy respectively. However, in general, very few schools (mostly less than 1.0 per cent) recorded difference in school effects greater than 0.125 logits (equivalent of about one school-term's work) in favour of either boys or girls.

Potential implications

The analyses and discussion presented in this chapter are interesting for at least two reasons.

First, the procedures described in this chapter could be employed to identify so-called 'extreme' schools for different groups of students, divided by such characteristics as prior achievement, socioeconomic status, race and migrant status. For purposes of identifying the extreme schools, it appears that either of the two approaches (split-school or varying effect) could be employed because they both yield very similar results. However, the split-school approach might be appropriate only where the distribution of the characteristic of interest is almost the same across the schools (for example, gender) and all the resulting sub-groups (boys, girls) are relatively large per school (at least 25 students, Raudenbush and Willms, 1995). Where the characteristic of interest is unlikely to be uniformly distributed across all the schools (for example, race) and some of the resulting sub-groups (ATSI, non-ATSI) are likely to be relatively small in some schools, it might be more appropriate to employ the varying effect approach. Nevertheless, there is need for further research on this issue.

Second, this chapter has demonstrated that school effects could be more relevant if expressed in terms of years of learning that a student spends at school. The same techniques could be employed to identify so-called 'extreme' schools when dealing with overall Types A, B and C school effects reported in preceding chapters. Obviously, expressing the school effects in terms of learning time lost or gained for the student could make the information more useful to educational administrators wishing to monitor school progress, and also to parents wishing to choose a school for their children. Moreover, it appears that based on this learning time lost or gained criterion, the differences between the most effective and the least effective schools are more apparent than when based on the criterion of variance between the schools. This is particularly important for the comparison of primary schools in South Australia because very small amounts of variance are left unexplained at the school-level, which makes the ranking of schools very unreliable and, thus, it is difficult to identify either a good school or a weak school.

12

Summary and Implications

The data sets for the current study were obtained from the Department for Education, Training and Employment (DETE) in South Australia. These data sets have been collected annually as student responses to Basic Skills Tests (BST) administered to Grades 3 and 5 students in government schools throughout South Australia since the inception of the Basic Skills Testing Program (BSTP) in 1995.

The major aims of this study are to:

(a) develop common scales for measuring achievement in the Basic Skills Tests across the Grades 3 and 5 primary school levels and across six testing occasions (1995 to 2000) in South Australia;

(b) examine the achievement levels of the Grades 3 and 5 students in the Basic Skills Tests in South Australia;

(c) examine changes in the numeracy and literacy achievement levels of Grade 5 students in South Australia;

(d) develop multilevel models of student-level and school-level factors influencing numeracy and literacy achievement of Grade 5 students in South Australia; and

(e) investigate the issues associated with measuring the value added components of the education provided in the South Australian primary schools, and how these measured components could be based on Grade 5 students' scores from the Basic Skill Tests.

Summary of the study

In order to achieve the above purposes the following decisions and procedures were undertaken.

Calibration: The Rasch model was selected for use in this study because the model allows item parameters to be estimated independently of the students sampled, and the student parameters to be estimated independently of the sample of items employed (Rasch, 1960; Keeves and Alagumalai, 1999). Based on the Rasch modelling procedures, it was necessary to examine the scaling characteristics of individual items

to determine whether or not each item employed contributed to the goal of effective measurement of student achievement in numeracy and literacy. It was also necessary to examine the response characteristics of each student who took the tests to determine whether or not there were misfitting persons, who needed to be excluded in the calibration process to avoid distorting the measurement scales (Wright, 1999).

Consequently, all the Grades 3 and 5 Basic Skill Tests from the six testing occasions (1995 to 2000) were calibrated separately using the Rasch model. In the calibration process, the fit statistics of the items and persons were examined for their conformities with the requirements of Rasch modelling. For this purpose, items having indicators of fit (INMS values) within the range 0.77 to 1.30 were considered to be appropriate (Adams and Khoo 1993), while persons having outfit t values within the range ±3.00 were considered appropriate. For the persons, this outfit t fit criterion (±3.00) employed to judge the fit of the Rasch model to the person in this study is more strict when compared with the range (±5.00) employed by Wright and Stone (1979) as well as Afrassa (2002).

Equating: It was necessary to equate all the Grades 3 and 5 tests from all the six testing occasions in order to form two common scales (one for numeracy and the other for literacy) upon which students' scores could be compared across grades and across testing occasions in South Australia. The main procedure selected to equate tests in this study was concurrent equating because research studies have shown that this procedure, when compared to alternative procedures, provided a more consistent and stronger measure of the two sets of items and persons being equated (Morrison and Fitzpatrick, 1992; Mohandas 1996).

The equating process was carried out in two main steps. Step 1 involved linking the Grades 3 and 5 tests from the same testing occasion in South Australia using the common items included in the tests. This step was undertaken in order to calculate the differences between (a) the achievement levels of the Grades 3 and 5 students from the same testing occasion, and (b) the average difficulty levels of the items included in the Grades 3 and 5 tests for the same testing occasion. For each subject, this first step yielded six separate scales (that is, one from each testing occasion). Step 2 involved linking all the Grades 3 and 5 tests from the six testing occasions in South Australia using some equating data sets obtained from the New South Wales Department of School Education. The differences calculated in Step 1 above were employed to check and strengthen the consistency of equating in the overall scales.

Scoring: After the successful equating of the tests from all the six testing occasions, the next step undertaken was to estimate Rasch scores of the students in numeracy and literacy. These scores were used as variables in subsequent analyses in this study.

For the purpose of scoring, the thresholds of equated test items were anchored and used to compute the scores for every student from the six testing occasions at Grades 3 and 5, counting omitted items and not-reached items as wrong. The scores of the students with perfect scores and those of students with zero scores were approximated manually because with Rasch modelling procedures scores of such students can not be estimated directly. In addition, a working rule was set not to calculate scores for students who did not respond to at least one item in the tests. This scoring procedure is consistent with the approach employed by the Australian Council for Educational Research (ACER) as well as the Basic Skills Tests developers in the Department of Education in New South Wales.

Development of achievement models: For the purposes of analysis, the hierarchical linear modelling technique was selected for use in this study after taking into consideration the multilevel structure of the available data (Raudenbush and Bryk,

2002). Based on this hierarchical linear modelling technique, and after taking into consideration the data sets available, three general types of models (Models X, Y and Z) of factors influencing student achievement in the Basic Skills Tests were developed.

Model-X was developed for teasing out the factors influencing numeracy and literacy achievement among Grades 3 and 5 students in South Australia by analysing the BSTP data collected from all the six testing occasions of interest in this study. This model was especially designed for estimating growth in achievement across the two grade levels.

Model-Y was developed for teasing out the factors influencing numeracy and literacy achievement among Grade 5 students in South Australia by analysing the transience data set, that is, the BSTP data on all the students who could be matched between Grade 3 and Grade 5. This model was especially designed for estimating the effects of transience on numeracy and literacy achievement of Grade 5 students in South Australia.

Model-Z was developed for teasing out the factors influencing numeracy and literacy achievement among Grade 5 students by analysing the non-transience data set, that is, the BSTP data on the students who could be matched and at Grade 5 they were in the same schools that they attended at Grade 3. This model was especially designed for estimating the effects of prior achievement on achievement of Grade 5 students in South Australia.

For each of the three general types of models noted above, two specific models were developed for each subject: a two-level model and a three-level model. For the two-level model, the hierarchical structure employed was students nested within schools while the hierarchical structure employed for the three-level models was students nested within schools, and schools in turn nested within testing occasions.

Development of school effects models: Based on the theoretical and statistical model for estimating school effects proposed by Willms and Raudenbush (1989), which itself is based on Carroll's model of school learning (Carroll, 1963), a general longitudinal three-level hierarchical model was proposed for estimating school effects in this study. The hierarchical structure employed in this longitudinal model was students nested within testing occasions, and testing occasions nested within schools. Within this structure, data on successive cohorts of students were used to map performance of the schools. Consequently, the structure employed in this study is very powerful because it separates a school performance index into a stable component and a change component, and also allows for the investigation of the factors influencing the performance slope for change over time. The stable component provided information about the average level of performance of the school over the study period, whereas the change component provided information on the trend in performance of the school over the time period under survey (Willms and Raudenbush, 1989).

From the general longitudinal analysis structure noted above, specific models were developed for measuring both stable and change for Types A, B and C school effects using students' scores in numeracy and literacy on the Grade 5 Basic Skills Tests as the outcome variables.

For each outcome variable, two models were developed and employed to estimate the stable and change components of each of the three types of school effects (or value added scores) noted above: a transience model and a non-transience model. The transience model used all the students who were matched (N=37,832) while the non-transience model used students who remained in the same school between Grades 3 and 5 (N=32,741).

In addition, the following two approaches were developed and employed to estimate Type A school effects for boys and girls in this study: split school approach and varying effect approach. Moreover, for Type A effects, two-level models were developed and employed to estimate school effects for each cohort of students.

Finally, the resulting value added scores were examined for their stability and their consistency across (i) transience and non-transience data sets (ii) subject areas, and (iii) types of school effects. Furthermore, for Type A effects, the resulting scores were also examined for consistency across testing occasions and for consistency across students' gender groups.

Answers to the research questions

In this section, the answers that were obtained to the research questions presented in Chapter 1 are outlined. The answers are presented under different sub-headings, which reflect the research issues addressed by the respective questions.

Calibration

(1) Is there adequate fit of the Rasch model to the Grades 3 and 5 items?

The results of Rasch analyses brought to light information regarding the Rasch scaling characteristics of the items in the BSTP. All 482 items in the Grades 3 and 5 Numeracy tests from all the six occasions have their indicators of fit (INMS values) within the desired range, 0.77 to 1.30. For the Literacy tests, only two items (Item 51 in the 1997 Grade 5 test [INMS=1.31], and Item 13 in the 2000 Grade 3 test [INMS=1.41]) of the 852 items have their fit values outside the stipulated range and were accordingly deleted. Thus, the Rasch model has adequate fit for virtually all of the items included in the 1995 to 2000 South Australian Basic Skills Tests.

(2) How do the average item difficulties of the Grades 3 and 5 tests compare across testing occasions?

The Rasch analyses also revealed information regarding the thresholds (or difficulties) of the test items included in the South Australian BSTP. Generally, the differences between the Grade 3 items mean and the Grade 5 items mean compare well from occasion to occasion, which indicates that the test developers did an excellent job in the development of the items and in the allocation of the items to either the Grade 3 or the Grade 5 tests.

Equating

(3) Can the numeracy items for 1995 to 2000 tests for Grades 3 and 5 be brought to a common scale?

(4) Can the literacy items for 1995 to 2000 tests for Grades 3 and 5 be brought to a common scale?

The analyses and discussion presented in Chapter 6 provide answers of 'yes' to both (3) and (4) above. Even so, those analyses and discussions revealed two technical points that must be taken into account when equating the South Australia BST data across occasions. First, under the current equating approach in which common persons data are used to link the tests across occasions, there are chances that the equating results could be distorted by students not trying as hard as possible in the trial tests (especially at Grade 5 level) and a practice effect (especially at Grade 3 level). Second, it is also likely that the equating results could be distorted where double links are used to equate the tests from the different testing occasions. However, it was not

immediately clear how these sources of errors could affect the equating of the Basic Skills Tests in South Australia, and consequently, there is need for further study to examine these equating issues. It would seem likely that the second issue is related to the differences in the standard deviations and the effects of these differences on item discriminations and scale factors employed in scaling. Moreover, the possibility arises from these analyses that there are some deficiencies in the development of the QUEST computer program, which lead to the program not doing what it is designed to do. Clearly, there is need for further research to investigate aspects of this problem.

Level of achievement

(5) Has the level of performance in numeracy (or literacy) at Grade 5 changed significantly over time?

The results of preliminary analyses reported in Chapter 6 indicate that, with only a few exceptions, the achievement in both numeracy and literacy at Grade 5 in South Australia has continued to increase since the inception of the program six years ago in 1995. In addition, those results of preliminary analyses indicate that the achievement in numeracy and literacy at the Grade 3 level has remained relatively constant. There are, however, some problems with the scaling procedures employed that must be taken into account in these comparisons.

For example, when considering only those students who had two data points (one at Grade 3 and the other at Grade 5), the results of multilevel analyses reported in subsequent chapters reveal that generally the achievement in both numeracy and literacy at Grade 5 in South Australia has declined significantly ($p<0.05$) over time. These declines in achievement remain significant even after student-level and school-level variables are included in the models. For example, metric regression coefficients for the linear trend variable OCC[29] that are obtained from the Type A effects models based on the transience and non-transience data sets for numeracy are -0.02 (t=-3.52) and -0.02 (t=-3.18) respectively, and for literacy are -0.10 (t=-15.70) and -0.10 (t=-15.20) respectively. For both subjects, these results follow closely the results obtained from the Types B and C models and also follow closely the results from the two-level analyses reported in Chapter 7. Hence, these results provide strong indications that the average levels of performance in numeracy and literacy at the Grade 5 level in South Australia have declined significantly ($p<0.05$) between 1995 and 2000, after allowance has been made for perturbations in scaling and for significant factors influencing student achievement.

(6) What is the average growth in numeracy and literacy achievement between Grades 3 and 5 levels?

The results of preliminary analyses reported in Chapter 6 indicate that the changes (or growth) in achievement between Grades 3 and 5 for both numeracy and literacy in South Australia are approximately half of a logit per year of school learning. In addition, the results that are obtained from the HLM analyses of the null as well as the final two- and three-level multilevel models are almost similar to the results obtained from the preliminary analyses (see Chapters 7 and 8). For example, the metric regression coefficients for the grade level variable YEARLEVL[30] that are obtained from the final two- and three-level models for numeracy are 0.51 (t=85.93) and 0.54 (t=82.44) respectively, and for literacy are 0.57 (t=88.16) and 0.55 (t=104.88)

[29] Coded: '1995/1997 Cohort' = 0, ..., '1998/2000 Cohort' = 3.
[30] Coded: 'Grade 3 Student' = 0, 'Grade 5 Student' = 2.

respectively. Thus, the average growth in numeracy and literacy achievement between Grades 3 and 5 is around one logit.

Factors influencing numeracy and literacy achievement

(7) What student-level factors influence numeracy (or literacy) achievement?

When considering the data on students who had two data points, the results of the two- and three-level analyses reported in Chapters 7 and 8 show that all seven student-level variables examined in this study have some significant ($p<0.05$) effects on achievement in literacy. These seven student-level variables are: Age of Student, Sex of Student, Racial Background, Speaking English at Home, Living in Australia (or Migrant Status), Transience, and Prior Achievement. All but one (Speaking English at Home) of these seven student-level variables also have significant influences on achievement in numeracy in both the two- and three-level models. Nevertheless, it should be noted that without the Prior Achievement variable in the models, Speaking English at Home also has a significant influence in numeracy. Indeed, this variable has a significant influence on achievement in numeracy when considering the data set that consisted of all the students who had participated in the BSTP (N = 144,346) because a prior achievement variable was not included in the hypothesized models (see Figures 5.2 and 5.5).

It is worth noting that, except for Prior Achievement (with effect sizes between 0.70 and 0.80), all the other student-level variables have very small effect sizes (mostly ≤ 0.10) especially in the final models. It is particularly important to note that, in analyses where a prior achievement variable is available for examination, this variable has by far the greatest magnitude of effect for both numeracy and literacy.

From these student-level results, the following conclusions can be drawn regarding achievement in numeracy and literacy of South Australian Grade 5 students when other factors are equal. Based on the effect sizes of variables, Prior Achievement (that is, achievement at Grade 3) is the most important predictor of student achievement in numeracy and literacy at Grade 5, with high achievers at Grade 3 being the ones most likely to achieve better at Grade 5. Nonetheless, younger students are likely to achieve better in both numeracy and literacy than their older counterparts, while students from a non-Aboriginal background are likely to achieve better in both subjects than students of an Aboriginal background.

In addition, students who always speak English at home are likely to achieve better in literacy (but not necessarily in numeracy) than students who rarely speak English at home. Furthermore, boys are likely to achieve better in numeracy than girls while girls are likely to achieve better in literacy than boys while, for both numeracy and literacy, students who are new to Australia are likely to achieve better than their counterparts who were born in Australia. Finally, students who remain in the same school over Grades 3 and 5 are likely to achieve better when compared to students who change schools. Indeed, it was estimated that, on the average, students who changed schools between Grades 3 and 5 were likely to lag behind their counterparts who remained in the same schools by about eight weeks and six weeks of school learning in numeracy and literacy respectively, after allowance had been made for other significant factors in the analyses undertaken.

Hence, the major difference between the student-level factors influencing achievement in numeracy and literacy is the role of student's gender: boys significantly outperform girls in numeracy, whereas girls significantly outperform boys in literacy. However, the magnitude of difference between boys and girls is trivial, as reflected by the very

small effect sizes of the variable SEX[31] (Sex of Student): -0.05 for numeracy, and 0.02 for literacy.

(8) What school-level factors influence numeracy (or literacy) achievement?

At the school-level, the two-level and three-level analyses reported in Chapters 7 and 8 show that of the 13 school-related variables examined in this study, four variables have significant ($p<0.05$) influences on achievement in both numeracy and literacy in a vast majority of the proposed models. These four variables are Proportion of School Cardholders, Locality of the School, Mobility and Absenteeism Rates. In some models, the variable School Size has significant influences on achievement in numeracy and literacy, and in some models, the Proportion of non-ATSI in the school has a significant influence for numeracy but not for literacy.

The magnitudes of effects of these school-level variables that have significant influences on achievement in numeracy and literacy are generally very small (with effect sizes mostly ≤ 0.10), especially when using data sets consisting of students who have two data points. Nevertheless, among these school-level variables, the socioeconomic status (Proportion of School Cardholders) and absenteeism variables have generally larger magnitudes of effects in a majority of the models for numeracy and literacy than the other significant variables.

When other variables are equal, these school-level results show the following findings regarding achievement in numeracy and literacy among Grade 5 students in South Australia. Students in schools with low proportions of School Cardholders (that is, with many higher socioeconomic status) students are likely to achieve better than their counterparts in schools with high proportions of School Cardholders. Students in urban schools or in schools located in (or near) Adelaide are likely to achieve better than students in rural schools or in schools located far from Adelaide. In addition, students in schools with low mobility rates are likely to achieve better in both numeracy and literacy than students in schools with high mobility rates while students in schools with low absenteeism rates are likely to achieve better than their counterparts in schools with high absenteeism rates.

(9) What cross-level interaction effects influence numeracy (or literacy) achievement?

The results of two- and three-level analyses reported in Chapters 7 and 8 also show that there are several cross-level interaction effects involving student-level variables and school-level variables (see results in Tables 7.8, 7.9 and 8.2 and 8.3). From these interaction effects, the following conclusions can be drawn.

(a) Locality of school has a greater influence on achievement in numeracy and literacy of ATSI students than of non-ATSI students, with ATSI students in rural schools being likely to achieve much lower than would be expected.

(b) Locality of school has a greater impact on achievement in numeracy of students who rarely (or never) speak English at home than of students who always speak English at home. As a result, students who rarely (or never) speak English at home are likely to achieve much lower in numeracy than what would be expected if they are in schools located in rural areas than if they were in schools located in urban areas.

(c) The students who always speak English at home are likely to achieve equally well in numeracy and literacy regardless of whether they are in schools with

[31] Coded: 'boy' = 0, 'girl' = 1.

high proportions of students born in Australia, or they are in schools with high proportions of students who are new to Australia. However, students who rarely (or never) speak English at home are likely to achieve better in numeracy and literacy if they are in schools with high proportions of students who are new to Australia than if they are in schools with high proportions of students born in Australia.

(d) The size of school has a greater impact on the achievement in numeracy and literacy of Grade 3 students than of Grade 5 students, with Grade 3 students in large schools achieving at a much lower level than would be expected.

(e) Transient students who move to schools that have high prior achievement scores in numeracy (or literacy) are likely to achieve better in numeracy and literacy than students who move to schools with average or low prior achievement scores in numeracy (or in literacy).

(f) Transient students who move into schools that have high proportions of younger students are likely to achieve better in numeracy and literacy than students who move into schools with high proportions of older students.

(g) Absenteeism in schools is likely to affect achievement in numeracy and literacy of the students with high prior achievement scores more than it affects the achievement of students with low prior achievement scores.

(h) Students who have high prior achievement scores in numeracy are likely to achieve better in numeracy if they are in schools with high proportions of boys than if they are in schools with high proportions of girls. On the other hand, students who have low prior achievement scores in numeracy are likely to achieve better in numeracy if they are in schools with high proportions of girls than if they are in schools with high proportions of boys.

(i) Students who have high prior achievement scores in numeracy are likely to achieve better in numeracy if they are in schools with high proportions of students born in Australia than if they are in schools with high proportions of new students to Australia. Conversely, students who have low prior achievement scores in numeracy are likely to achieve better in numeracy if they are in schools with high proportions of students who are new to Australia than if they are in schools with high proportions of students born in Australia.

In addition, for both numeracy and literacy, the results of analyses reported in Chapter 10 (Table 10.3) show that there are significant interaction effects between the linear trend variable (OCC) and the school size variable (SSIZE_2). From these interaction effects, it can be concluded that during the earlier testing occasions, students in large schools were likely to achieve better in numeracy and in literacy than their counterparts in small schools. However, during the later testing occasions, students in small schools were likely to achieve better in both subjects than students in large schools.

Variance partitioning and variance explained

(10) What amounts of variance are available at the student-level, school-level and occasion-level?

The results of variance partitioning based on two-level and three-level analyses are presented in Chapters 7 (Table 7.1) and 8 (Table 8.1) respectively. Based on the transience data set and on a two-level analysis, the results of variance partitioning show that 81.7 and 18.3 per cent of the variance of student numeracy scores are at the student and school levels respectively. And based on a three-level analysis the

percentages are 81.8, 18.1 and 0.2 for student, school and occasion levels respectively. The corresponding percentages based on the non-transience data set from a two-level analysis are 82.1 and 17.9 for student and school levels respectively, and from a three-level analysis are 82.2, 17.6 and 0.2 for student, school and occasion levels respectively.

Based on the transience data set, the corresponding percentages for student literacy score from a two-level analysis are 81.8 and 18.2 for student and school levels respectively, and the corresponding percentages from a three-level analysis are 81.9, 16.3 and 1.8 for student, school and occasion levels respectively. For the non-transience data set, these percentages from a two-level analysis are 81.7 and 18.3 for student and school respectively, and from a three-level analysis 81.8, 16.3 and 1.9 for student, school and occasion respectively. These percentages of variance of student scores at the various levels of the hierarchy are the maximum amounts of variance available at those levels that could be explained in subsequent analyses.

The results of variance partitioning based on the transience data set overwhelmingly agree with the results based on the non-transience data set, and the results of variance partitioning for numeracy follow very closely the results for literacy. In addition, for both outcome measures, the variation of student scores at the student and school levels based on the two-level analysis overwhelmingly agree with the variation at these levels based on the three-level analysis. Thus, as far as the amount of variance available to be explained at the student and school levels are concerned, it does not seem to matter markedly whether a two-level or three-level analysis is employed.

From these variance partitioning results, it can be concluded that in South Australia, the variance between students within schools in terms of their achievement in numeracy and literacy at Grade 5 is roughly around four times greater than the variance in performance between schools. That is, there is huge variability within the schools when compared to the variability between schools. In addition, the results from the three-level analyses reveal that very small (less than 2.0 per cent) of the variation between student scores can be attributed to the testing occasions.

(11) What percentages of variance in student scores in numeracy and literacy at Grade 5 are explained by Prior Achievement (that is, achievement at Grade 3) alone?

The results of the amounts of variance explained by Prior Achievement for numeracy and literacy are reported in Chapter 7 (Table 7.4) based on a two-level analysis. For numeracy, these results show that Prior Achievement explained 46.2 and 46.9 per cent of the total variance available in the models based on in the transience and non-transience data sets respectively. For literacy, the percentages of total variance explained by Prior Achievement alone were 55.7 and 56.6 based on the transience and non-transience data sets respectively.

Based on either the transience or non-transience data sets, Prior Achievement explained about two fifths (40 per cent) and about a half (50 per cent) of the amount of variance available at the student-level for numeracy and literacy respectively. Furthermore, for both subjects, Prior Achievement explained approximately two thirds (60 per cent) of the amount of variance available at the school-level based on either the transience or the non-transience data sets.

In general, these percentages of the amount of variance explained by Prior Achievement at each level of the hierarchy do not differ markedly from the percentages of variance explained in the final two-level models. Consequently, it can be concluded that as far as the amounts of variance explained are concerned, Prior

Achievement is by far the most important predictor of student achievement at Grade 5 among the predictors examined in this study.

(12) What percentages of variance in student scores in numeracy and literacy at Grade 5 do the predictor variables in the final two-level and three-level models explain?

For a two-level analysis, the results of the amounts of variance explained by the predictor variables in the final model for numeracy and literacy are presented in Table 7.10 while Table 8.4 displays the corresponding information for a three-level analysis. For numeracy, these results show that predictor variables included in the final two-level model explained 49.4 and 49.7 per cent of the total variance based on the transience and non-transience data sets respectively, and in the final three-level model these percentages are 49.5 and 49.8 respectively. For literacy, the corresponding percentages for the transience and non-transience data sets are 58.4 and 58.5 respectively for the final two-level model, and are 56.9 and 57.0 respectively for the final three-level model.

For both outcome measures, the total amounts of variance that are explained by the predictors in the transience model followed very closely the amounts that are explained in the non-transience model. Importantly, the total amounts of variance that are explained based on a two-level analysis are basically the same as the amounts that are explained based on a three-level analysis. Therefore, the three-level analysis offers no added advantage as far as the amounts of variance explained in the final models are concerned.

In addition, the results reported in Chapters 7 and 8 show that the percentage of variance explained at the school-level for numeracy and literacy are over 70 per cent of the total variance available at the school-level, regardless of the type of analysis employed and regardless of the data set analyzed. As a consequence, small amounts (around five per cent) of variance that are available at the school-level are left unexplained for numeracy as well as for literacy.

School effects

The results of the amounts of variance explained by the predictor variables in the longitudinal model for estimation of Types A and B effects are presented in Chapter 9 (Tables 9.7 and 9.8), and the corresponding information for Type C effects is presented in Chapter 10 (Table 10.5). These results provide answers to research questions 13 to 15.

(13) What percentages of variance in student scores in numeracy and literacy at Grade 5 are explained in the models employed to estimate school effects?

For numeracy, the percentages of variance explained based on the transience and non-transience data sets are (58.5 and 56.7) for Type A effects, (60.3 and 58.1) for Type B effects, and (60.3 and 58.2) for Type C effects respectively. For literacy, the corresponding percentages for transience and non-transience data sets are (64.7 and 64.2), (65.5 and 64.8), and (65.5 and 64.8) for Types A, B and C effects respectively.

For both subjects, these results show that, regardless of the data set (transience or non-transience), and regardless of type of school effect being estimated, roughly 60 per cent of the amounts of variance available in the models are explained. Furthermore, these results show that for both outcome measures, the amounts of variance explained in the Type B effects models are basically the same as the amounts that are explained in the corresponding Type C effects models.

By and large, based on the three-level longitudinal design employed to estimate school effects, higher amounts of variance are explained in the final models compared to the amounts that are explained based on the two- and three-level models employed to tease out factors influencing student achievement. Therefore, if the total amounts of variance explained are to be used as a measure of how well a model fits the data, then the models under the longitudinal structure are better models compared to two-level and three-level models employed to tease out factors influencing student achievement in Chapters 7 and 8 respectively.

(14) What percentages of variance in student scores in numeracy and literacy at Grade 5 are left unexplained at the student-level in the models employed to estimate school effects?

For numeracy, the percentages of variance left unexplained at the student-level based on the transience and non-transience data sets are (36.3 and 37.6), (36.2 and 37.6), and (36.2 and 37.6) for Types A, B and C effects respectively. These percentages for literacy are (31.1 and 31.4), (31.2 and 31.5), and (31.2 and 31.5) for Types A, B and C effects respectively.

Within the same outcome measure, these results show that, regardless of the data set (transience or non-transience) and type of school effect being estimated, the percentages of variance left unexplained at the student-level remain almost the same. These results also show that the percentages of variance left unexplained at the student-level in the numeracy models (around 36 to 38) are generally larger compared with the percentages of variance left unexplained in the literacy models (around 31 to 32). Nonetheless, for both outcome measures, these percentages of variance left unexplained at the student-level are arguably large, and therefore, it can be concluded that there are other important student-level (and probably class-level) factors influencing student achievement that are not included in the models employed in this study.

(15) What percentages of variance in student scores in numeracy and literacy at Grade 5 are left unexplained at the school-level in the models employed to estimate school effects?

For numeracy, the percentages of variance left unexplained at the school-level based on the transience and non-transience data sets are (2.9 and 2.9), (1.1 and 1.4), and (1.1 and 1.4) for Types A, B and C effects respectively. The percentages for literacy are (1.9 and 1.6), (1.0 and 1.0), and (1.0 and 1.0) for Types A, B and C effects respectively.

For both numeracy and literacy, these results show that regardless of the type of effect being estimated, the percentages of variance left unexplained at the school-level are very small. In particular, the percentages of variance left unexplained are extremely small (around one per cent) in Types B and C models, and therefore, it can be concluded that almost all the important school-level factors influencing student achievement are included in these models.

Importantly, compared to the amount of variance explained at the school-level by student-level variables alone in Type A effects models, the school-level variables included in the Types B and C models explain only very small amounts of the variance at the school-level. Thus, most of the differences between schools in their performance in the Basic Skills Tests at Grade 5 can be explained by the differences in their student intake characteristics, measured at Grade 3.

(16) How reliably are the school effects estimated?

The school-level reliability estimates from Types A and B effects models are presented in Chapter 9 (Table 9.1), and the corresponding information from Type C effects models are presented in Chapter 10 (Table 10.1). From these results, it can be concluded that substantial percentages of the variance between estimates of the average school performance (mostly more than 50 per cent) and between estimates of change in school performance (more than 80 per cent) are random fluctuations that could be associated with measurement and sampling errors. However, under the multilevel estimation procedure employed in this study, the school effects are automatically adjusted in a process called 'shrinkage' to cater for sampling and measurement errors in order to present a more accurate picture of the variation between schools (see Willms 1992; p.42). Nevertheless, these results show that, at Grade 5 primary school level in South Australia, it is difficult to detect differences between schools in their average performance (stable effects) and in their improvement or deterioration (change effects) in performance over time.

Notwithstanding the above findings, the following statements can generally be made regarding the reliability estimates of the various indices of school effectiveness computed in this study.

(a) For both outcome measures, the stable components of Types A, B and C school effects are estimated far more reliably (λ = 0.37 to 0.60) than their corresponding change components of school effects (λ = 0.13 to 0.19). Thus, if the performance of a primary school in South Australia is to be judged on the value the school has added to the achievement of its Grade 5 students, more faith should be placed on stable school effect indices than on the change school effect indices.

(b) For the stable component of school effects and for both numeracy and literacy, Type A effects (λ = 0.46 to 0.64) are estimated more reliably than are the Type B effects (λ = 0.37 to 0.45), and Type C effects (λ = 0.37 to 0.44) are estimated almost as reliably as Type B effects. Hence, at Grade 5 primary school level in South Australia, more reliance could be placed on Type A effect indices than on Type B (or Type C) effect indices.

(c) Generally, the stable school effects for numeracy (λ = 0.43 to 0.64) are estimated slightly more reliably than for literacy (λ = 0.37 to 0.53). Therefore, at Grade 5 primary school level in South Australia, slightly more reliance could be placed on the school effect indices for numeracy than on school effect indices for literacy.

(17) Can a stability index be calculated to compare the stability of the various types of school effects over time?

The stability ratio criterion that was developed by Willms and Raudenbush (1989) was employed in this study to compare the stability of the various types of school effects over time. The stability ratios for Types A and B effects are presented in Chapter 9 (Tables 9.5 and 9.6 respectively), and in Chapter 10 (Table 10.4) for Type C effects. From these stability ratios, the following conclusions can be drawn.

(a) Type A effects (stability ratio = 7.17 to 12.20) are more stable over time than either Type B or Type C effects. And Type C effects (stability ratio = 5.12 to 5.84) are marginally more stable over time than Type B effects (stability ratio = 4.73 to 5.17).

(b) School effects for numeracy (stability ratio = 4.94 to 12.20) are marginally more stable over time than for literacy (stability ratio = 4.73 to 8.10).

(18) Based on value added scores, is the rank order of the schools, using all the students who could be matched, greatly different from the rank order of the schools using only those students who could be matched in the same school?

Within the same outcome measure and the same type of school effect, the results of analyses reported in Chapter 9 (Tables 9.9 and 9.13) and Chapter 10 (Table 10.6) overwhelmingly show the existence of extremely strong correlations between school effects based on the transience and the non-transience data sets ($r \geq 0.94$). Thus, it can be concluded that based on the same type of school effects, the rank order of schools obtained using all the students who could be matched and the rank order of schools based on the students who remained in the same school do not differ markedly.

(19) Are schools that are identified as relatively effective based on one type of school effect also identified as relatively effective based on a different type of school effect?

The correlations between Type A and Type B effects are presented in Chapter 9 (Table 9.10), and those between Type B and Type C effects are presented in Chapter 10 (Table 10.8). Within the same subject (i.e. numeracy or literacy), these results show very strong to extremely strong correlations (0.78 to 0.90) between the stable Type A and Type B effects, and very strong to unity correlations (0.94 to 1.00) between the stable Type B and Type C effects. Similarly, extremely strong to near unity correlations are evident between change Types A and B effects (0.85 to 0.98), and between change Types B and C effects (0.92 to 0.99). From these results, the following conclusions can be drawn.

(a) Within the same subject, a vast majority of the primary schools that perform well based on Type A effects also perform well based on Type B effects, and vice versa.

(b) Within the same subject, basically all the primary schools that perform well based on Type B effects also perform well based on Type C effects, and vice versa. Thus, for purposes of ranking the primary schools in South Australia, it is unnecessary to compute Type B effects because the ranks assigned to the schools would basically be the same as the ranks assigned to the schools based on Type C effects.

(20) Do schools that show more than expected average levels of performance also show more than expected increases in performance over time?

For Types A and B effects, the correlations between the stable and change components of school effects are presented in Chapter 9 (Table 9.11), and these correlations for Type C effects are presented in Chapter 10 (Table 10.7).

For numeracy, the results show small to medium but positive correlations between the stable effects and change effects for Type A effects (r = 0.27 to 0.31), for Type B effects (r = 0.42 to 0.43), and for Type C effects (r = 0.48 to 0.50). However, for literacy, the results show very strong to extremely strong but negative correlations between the stable effects and change effects for Type A effects (r = -0.76 to -0.84), for Type B effects (r = -0.68 to -0.77), and for Type C effects (r = -0.71 to -0.80). From these results, the following conclusions can be made regarding the performance of the primary school in South Australia based on the value the schools add to the achievement of their students at Grade 5.

(a) For numeracy, a considerable number of the primary schools that show more than expected average performance are likely to show more than expected increase in performance over time, and vice versa.

(b) For literacy, a vast majority of the primary schools that show more than expected average performance are highly likely to show less than expected increase in performance over time, and vice versa

(21) Are schools that are relatively effective in numeracy also relatively effective in literacy?

The correlations between school effects for numeracy and for literacy are presented in Chapter 9 (Table 9.9) for Types A and B effects, and in Chapter 10 (Table 10.6) for Type C effects. For stable school effects, these results show very strong correlations (0.67 to 0.71) among Type A effects, large correlations (0.50 to 0.56) among Type B effects, and medium to large correlations (0.49 to 0.55) among Type C effects. For change school effects, the correlations between school effects on numeracy and on literacy are small among Type A effects (0.13 to 0.18), among Type B effects (0.24 to 0.28), and among Type C effects (0.18 to 0.23). From these results, the following conclusions can be drawn regarding the consistency of school effects across the two subjects included in the BSTP.

(a) Generally, the stable Types A, B and C school effects are to some extent consistent across the two outcome measures included in the BSTP. That is, based on the same type of stable school effect, a considerable number of schools that show more than expected average levels of performance in numeracy are also likely to show more than expected average levels of performance in literacy, and vice versa. Nevertheless, when compared to the stable Types B and C school effects, the stable Type A effects are more consistent across the two subjects.

(b) The change Types A, B and C school effects are hardly consistent across the two outcome measures. That is, based on the same type of change school effect, only a small number of schools that show more than expected increase in performance over time in numeracy are likely to show more than expected increase in performance over time in literacy, and vice versa.

(22) Are schools that are relatively effective for one cohort of students also relatively effective for other cohorts of students?

In order to answer this question, two-level models that treated each school as a separate entity for each of the four cohorts of students were employed to estimate four Type A effects indices for each school: one for each cohort of students. The correlations among these Type A effects indices are presented in Chapter 9 (Table 9.13).

For both outcome measures, the correlations between these Type A school effects are positive but they are generally small though a few are medium (0.16 to 0.40). From these results, it can be concluded that only a small number of schools that show more than expected average performance with one cohort of students are likely to show more than expected average performance with another cohort of students. In other words, at the Grade 5 primary school level in South Australia, only a small number of schools that are relatively effective for one cohort of students in numeracy (or in literacy) are likely to be relatively effective for another cohort of students, and vice versa.

(23) Are schools that are relatively effective in numeracy for boys also relatively effective for girls?

(24) Are schools that are relatively effective in literacy for girls also relatively effective for boys?

In order to answer these two questions, two approaches were employed to estimate the stable Type A school effects for boys and girls: split-school approach and varying effect approach. The results of these analyses are reported in Chapter 11. Within the same gender of students, the correlations between the stable Type A schools effects obtained using the varying effect approach and those obtained using split-school approach are extremely strong (r = 0.89 to 0.94). Thus, it can be concluded that the rank order of the schools based on stable Type A effects obtained using these two approaches do not differ markedly.

Within the same outcome measure, the correlations between the stable Type A school effects for boys and girls are very strong based on the split school approach (r = 0.71 to 0.77), and extremely strong based on the varying effect approach ($r \geq 0.95$). Furthermore, within the same data set and based on the varying effect approach, the correlations between the stable Type A effects for boys and girls are close to unity ($r \geq$ 0.98) for both numeracy and literacy. From these results, it can be concluded that a vast majority of schools that record more than expected average performance in numeracy (or literacy) for boys also record more than expected average performance for girls in numeracy (or literacy), and vice versa.

The above conclusions having been drawn, the following should nevertheless be emphasized. Based on 0.13 logits as an indication of the amount of learning done in numeracy and literacy within one school-term, there was clear evidence that some schools that are effective for one gender of students could be substantially ineffective for the other gender of students.

Important issues, findings and implications

This study raises a number of issues and provides several findings that have potential implication for theory, policy, practice, and further studies on equating of the Basic Skills Tests, factors influencing student achievement and measurement of school effects in primary schools in South Australia. The accounts of these issues, findings and implications are provided in concluding sections of Chapters 6, 8, 9, 10 and 11. The purpose of the three sub-sections that follow is to provide summaries of the most important of these issues, findings and implications. The first sub-section looks at issues that could be of concern in the equating of the Basic Skills Tests across occasions while the second sub-section focuses on factors influencing student achievement. The last three sub-sections focus on issues, findings and implications related to school effects in South Australia.

Concerning equating

Key finding:

- *Use of common person and double links may distort equating results.*

On the whole, the analyses and discussion presented in Chapter 6 raises issues regarding the appropriateness of two approaches that are currently employed to equate BST data across occasions: use of common persons, and use of double links.

Under the common persons equating approach, the analyses reported in Chapter 6 provide some evidence that the equating results could be distorted by errors associated

with students not trying hard in the trial tests (especially at Grade 5 level) and a practice effect (especially at Grade 3 level). There are also concerns associated with differences in the distribution of the outcome variable in the equating sample and the main sample (Linacre and Wright, 1989; DeMars 2001 & 2002), and concerns associated with the curriculum differences between the equating State (New South Wales) and the main State (South Australia) giving rise to differences in variance.

In addition, the analyses reported in Chapter 6 provides some evidence that the equating results could be distorted by errors associated with use of double links to equate tests from different testing occasions.

In order to avoid the errors noted, it is recommended (in Chapter 6) that a preferred procedure would be for the test developers to include some common items in the tests across occasions so as to allow more accurate linking over time. In addition, it is recommended that the test developers would need to keep the tests secure to avoid students on future occasions obtaining access to the common items. However, it is not exactly clear how the sources of errors noted influence the equating of the Basic Skills Tests in South Australia, and consequently, there is need for further study to examine these equating issues.

Concerning factors influencing student achievement

Key findings:

- *Prior achievement, age, gender, racial background, migrant status, transience and English spoken at home are among the important individual-level predictors of student performance in the BSTP at Grade 5 in South Australia.*
- *Average socioeconomic status, location, mobility and absenteeism rates are among the important group-level predictors of student performance in the BSTP at Grade 5 in South Australia.*

The results of multilevel analyses presented in this study are interesting especially because they are based on data from several cohorts of students, modelled simultaneously. The results are also interesting because they show how student-level and school-level factors influence student achievement when considered simultaneously. Thus, parents, teachers and policy makers need to be aware of the role played by the factors identified here in student achievement in numeracy and literacy.

However, in the models developed in this study, large amounts of variance are left unexplained at the student-level, which show that there are other factors not included in the models that contribute to the variability in students' achievement in numeracy and literacy. Thus, the models developed in this study are just an initial step towards the development of more comprehensive models that would promote a greater understanding of the factors influencing student achievement in the BSTP in South Australia. Issues related to the development of such models are described in the sub-section that follows.

Concerning school effects

Key issue:

- *Why is the relationship between the stable school effects and the change school effects for numeracy found to be completely different from that of literacy?*

This study found that a considerable number of primary schools that show more than expected average performance in numeracy also show more than expected increase in performance in numeracy over time, and vice versa. On the contrary, for literacy, this

study found that a large number of primary schools that show more than expected average performance also show less than expected increase in performance over time, and vice versa.

Several attempts are made to provide an explanation to the above finding (see Appendix 14.6) but it remains unclear why the relationship between the stable school effects and the change school effects for numeracy is different from that of literacy in South Australia. Thus, there is a clear need for further study to investigate this issue.

Key finding:

- *In relative terms, primary schools in South Australia are not consistently effective (or ineffective) across time*

For both outcome measures, the correlations between the Type A school effects for each cohort of students (that is, estimated using data on each of the four cohorts of students), are positive but they are generally small though some of the correlations are of medium size (r = 0.13 to 0.40).

Hence, it is logical to question the appropriateness of computing school performance indicators using students' scores obtained from the Basic Skills Tests and using a single cohort of students given that such indicators could end up being used to compare or rank schools.

Clearly, the performance of the primary schools in South Australia cannot be judged reliably based on single cohort of students, and hence the strength of the longitudinal structure employed in this study.

Key finding:

- *Variance left unexplained at the school-level is very small (about one to three per cent)*

Without doubt, the discussions and analyses presented in this study show that after taking into account student background characteristics and achievement in the Basic Skills Tests at Grade 3, a very small amount of the variance available in the Grade 5 scores is left unexplained at the school-level. That is, most of the differences between schools in their performance on the Basic Skills Tests at Grade 5 can be explained by the differences between schools in their student intake characteristics, measured at Grade 3. Raudenbush and Willms (1995) argue that, for parents, if the variation between schools was very small, there would be no consequences for the expected achievement of a child when choosing among a set of schools; whereas, if the variation were large, such choices would be of crucial importance. For policy makers and administrators, Raudenbush and Willms, (1995; p. 315) argue that the magnitude of the variation between schools is important because it is an indicator "of the extent of inequality produced by the schooling system". Thus, these results have substantial implications for research into school effects of primary schools in South Australia, especially if the scores from the Basic Skills Tests are to be used as inputs for computation of the indicators of school performance across the two grade levels.

Although school effects can influence within-school variation by interacting with student background (Raudenbush and Willms, 1995), it is the amount of variance left unexplained that is ultimately important in the stability of the ranks assigned to schools based on either Type A or Type B or Type C school effects. Thus, despite the improvement possible in the stability of the ranks assigned to schools with the longitudinal models employed in this study, it must be asked whether it is appropriate to rank schools based on the small variance left unexplained. If so, how accurate or how reliable would such comparison or ranking of schools be? Parents in South

Australia choosing a school for their children would be interested in the Type A effect indicator, while the general public in South Australia could use the Types B and C indicators to hold schools accountable for their performance. But how reliable or useful would the information provided by these indicators be to the parents or to those in a position to hold schools accountable given that only very small amounts of variance are left unexplained at the school-level?

However, it should be remembered that although the variability between schools is small, this does not imply that it is unimportant or trivial (see Peaker 1975; p.140). Neither does it imply that schooling in South Australia does not result in an increase in student achievement. Indeed, this study show that even after controlling for factors influencing student achievement, Grade 5 students are better achievers than Grade 3 students by about one logit for both numeracy and literacy.

Thus, the argument is that there are very small differences between the contributions that the primary schools in South Australia make to the increase in their students' achievement across the two grades. That is, when everything else is equal, the school attended by a student between Grades 3 and 5 is of extremely little consequence to the student level of achievement in numeracy and literacy at Grade 5. Ultimately, the argument is entirely on the stability (and therefore, the usefulness) of the ranks assigned to schools based on the small amounts of variance left unexplained rather than the statistical significance of the variance left unexplained at the school-level for whatever type of school effects.

Although there are no simple answers to the above questions, it is obvious that few differences exist between the primary schools in South Australia that would warrant any comparison or ranking being made using the scores from the Basic Skills Tests. Consequently, the school performance indicators or ranks computed using the scores from the Basic Skills Tests need to be interpreted with a great caution.

Finally, if the small variance left unexplained at the school-level is borne in mind, it becomes clear why primary schools in South Australia appear not to be consistently effective (or ineffective) over time (see the previous key finding): the ranking assigned to schools is unstable because of the small residual variance. Thus, the rankings vary considerably from year to year.

Key finding:

- *Variance left unexplained at the student-level is incredibly large when compared to variance left unexplained at the school-level.*

This study show that substantial variability between the students is left unexplained (about 30 to 40 per cent) while almost all variability between the schools is explained in the longitudinal models employed to estimate school effects. Because of the small amount of variance left unexplained at the school-level (about one to three per cent), it is difficult to identify reliably weak schools, and it is also difficult to identify reliably good schools. Thus, it is argued in Chapter 10 that this finding has potential implications for policy in the funding of the primary schools in South Australia. It is argued that the practice of identifying weak schools and providing them with funds would not seem appropriate. Consequently, it is recommended that the government should focus on identifying weak students within schools and providing them with remedial programs to help them gain the required skills.

The above finding also has important potential implications for further studies aimed at investigating the causes of variability within schools and development of school effectiveness models that would best explain the situation within schools in South Australia. These implications are outlined in the next two sub-sections.

Examination of more student-level factors

It has been mentioned in an earlier sub-section that the data available for the current study lack some variables that might have explained some of the variability between the students. Certain important student-level variables that are not available for examination in this study include socioeconomic status, homework, grade repetition and student absenteeism. It is highly likely that inclusion of these variables at the student-level (especially SES) could bring down the amount of variance left unexplained at that level. Thus, there is a clear need for a thorough survey in order to establish what other items need to be included in the student questionnaire employed in the BSTP. Such a survey would provide information needed to construct student-level variables that could explain some of the variation between students.

Notwithstanding these recommendations, it is unlikely that inclusion of other important predictors of student achievement at the student-level (such as a SES variable) would lower noticeably the variance left unexplained at the student-level in South Australia. For example, as is argued in Chapter 10, it is reasonable to expect in South Australia that most of the variance associated with SES is entangled with the other student-level variables (such as Racial Background and Prior Achievement) and, therefore, has already been catered for in the model. Moreover, in Australia, results from the Longitudinal Survey of Australian Youth (LSAY) study showed that, "socioeconomic status at the individual-level has minimal influence on academic achievement, but at the school-level it has much greater influence" (Rothman and McMillan, forthcoming; p.25).

Thus, it would appear that, within the general hierarchical structures employed in this study (students nested within schools), it is unlikely that the variance left unexplained at the student-level would be decreased markedly by merely including other factors into the models. Consequently, there is a clear need for careful consideration of what could be happening within schools. This issue is dealt with next.

Future of school effectiveness research in South Australia

Results from studies that have taken into account differences among classes within schools indicate that those differences are usually as large as the differences between schools (e.g., Creemers and Reezigt, 1996; Einsiedler and Treinies, 1997; Fitz-Gibbon, 1991 & 1997; Kyriakides et al., 2000). More importantly, evidence is now available to show that inclusion of a class-level in multilevel analyses might help to disentangle the amounts of variances available at the class-level from the amounts of variances available at the student-level without necessarily changing the amounts of variances available and explained at the school-level. That is, when the class-level is included in the analysis, some of the variation between students (not between schools) can be attributed to the differences between classes the students belong to. Judging from the amounts of variance left unexplained in the school and student levels in this study, the same could be the situation in South Australia. In other words, the inclusion of a class-level of analysis in this study, could have most likely helped to bring down the variance left unexplained at the student-level without necessarily changing the amounts available and explained at the school-level.

The evidence mentioned above was obtained from my recent study looking at mathematics and reading achievement of 36,476 Grade 5 students in 7,221 classes in 3,635 schools in another country. In that study, it was also found that class-level factors such as teacher's years of professional training, teacher's knowledge and skills on the subject matter, and classroom resources and materials had significant ($p<0.05$) influences on student achievement in mathematics and reading. Moreover, in terms of effect sizes, teacher's knowledge or skills on the subject matter was found to be by far

the most important predictor of student achievement, with students who were taught by teachers who were more knowledgeable or skillful in the subject matter achieving markedly better.

From a school effectiveness perspective in South Australia, inclusion of a class-level on the analysis has the potential of providing a better picture of the situation within the primary schools compared to the image provided in this study. Evidently, there is a clear need for a further study to develop school effectiveness models that are the most appropriate for explaining the within schools variability in student achievement in South Australia and which include a class-level in the hierarchy. For such models to be achieved, it would require the development of a thorough data collection instrument aimed at capturing the pieces of information needed to construct the important variables at least at the student, class and school levels. Creemers' (1994) comprehensive model of school effectiveness, which is an extension of Carroll's model of school learning (Carroll, 1963), would be good starting point for the development of such a data collection instrument. Indeed, Creemers' model was employed successfully by Kyriakides et al. (2000) to develop data collection instruments in their school effectiveness work with primary schools in Cyprus.

At the class-level, Creemers (1994) in his model has specifically highlighted the importance of teacher and group attributes in student achievement, and careful examinations of these attributes would definitely promote a greater understanding of their influence on achievement in numeracy and literacy among Grade 5 students in South Australia.

Final words

Key issue:

- *What does value added involve?*

The inception of the BSTP in South Australia in 1995 marked the beginning of ongoing heated debate between proponents and opponents of the program. A majority of those opposed to the BSTP are mainly teachers, while the advocates of the program are mainly parents and politicians within the State Government. The critics of the program mainly argue that the program is unnecessary because the information that is obtained from the program about levels of achievement of the individual student, is in no way superior to what the teachers can gather when teaching, based on their professional training and experiences. On the other side of the debate, proponents claim that the program provides useful feedback to parents, teachers and educational administrators and that this feedback is necessary if weaker students are to be identified and assisted to acquire the necessary basic skills of numeracy and literacy.

So far there has been no attempt to rank or to publish the performance of the schools based on the their students' scores from the BSTP. However, proponents of the program have been quoted in a wide range of print and electronic media as having claimed that the results from the program have shown that the levels of achievement of successive cohorts of students have continued to increase since the inception of the program. And it has been claimed that the results from the BSTP show that schools have improved in their performance since the inception of the program in 1995. As would be expected, these claims have brought a new twist to the debate: that of the potential role of the BSTP as an instrument for assessing the performance of the primary schools in South Australia. Of course, this is the main reason teachers have been opposed to the program. That is, the results from the program could be used to rank schools and somehow to hold them and their teachers accountable for their students' levels of achievement in numeracy and literacy.

The purpose of this study was neither to dispute nor support one side or the other in the debate. Nor was the purpose of this study to develop new methods of school assessment or dispute the existing methods, but rather to investigate using the existing methods how school effects, could be measured based on students' scores from the BSTP. Specifically, this study sought to bring some research to the debate (especially with respect to performance of the schools) by developing a general model upon which school effects could be estimated, and towards clarifying the idea of what 'value added' involves.

This study has argued that, if schools were to be assessed in terms of the value added to students' achievement over a two-year period, then it would be necessary to allow for the performance of the students, before the commencement of the period under review. Critics of this study are likely to argue that in measuring school performance across Grades 3 and 5, it is inappropriate to consider Grade 3 score as accounting for prior-achievement. This is because the student's achievement at the Grade 3 level has the school's contribution already embedded in it from Grades 1 and 2. However, if the focus of the analyses were to measure the value added by a school to student achievement in the Basic Skills across the two grade levels, then adjustment for prior achievement at Grade 3 is appropriate. Surely, two years is a substantial time period for schools to have made a further ample impact on the achievement of their students. Thus, the question asked in this study is: how much has the school contributed to the student's achievement since Grade 3? And thus, have schools changed in their performance over time? It should be borne in mind that prior achievement variables (Grade 3 scores) and the outcome variables (Grade 5 scores) are measured on the same scale in this study, making it meaningful and fair to compare the achievement of the students across the two grades.

Hence, based on the meaning associated with the term value added, three types of school effects are identified in this study: Type A, Type B and Type C. For purposes of this study, these three types of school effects are defined as follows:

Type A: the contribution of a school to the increase in student achievement at Grade 5 after controlling for effects of student background characteristics and achievement at Grade 3;

Type B: the contribution of a school to the increase in student achievement at Grade 5 after controlling for effects of student background characteristics and achievement at Grade 3 together with the effects of the average school context;

Type C: the contribution of a school to the increase in student achievement at Grade 5 after controlling for effects of student background characteristics and achievement at Grade 3 together with the effects of average school context and school characteristics (for this study, school size and school location, namely; urban/rural).

The meanings associated with Type A and Type B effects in this study are the same as the meanings associated with these effects by Raudenbush and Willms (1995) and also by Harker and Nash (1996). Thus, parents wishing to choose a school for their child would be interested in Type A effects, while administrators wanting to hold a school accountable for its performance would be interested in Type B effects. For the South Australian situation, the Type C effect is a refined form of Type B effect and is aimed at capturing the contribution of the school to the increase in student achievement, free from the influence of funds provided to the schools by the government according to school size and school location.

Each of the three types of school effects defined in this study consist of two components: a stable component that shows the average performance of the school over the study period, and a change component that shows whether the performance of a school has improved or deteriorated over the study period.

Despite identifing the above types of school effects and illustrating how these effects could be estimated, this study shows that, within the South Australian situation, it is very difficult to identify effective or ineffective schools because the amount of variance left unexplained at the school-level is very small. As a solution to this problem, this study demonstrates that it is more meaningful to identify effective or ineffective schools when the school effects are expressed in terms of years of learning that a student spends at school. Even so, the study argues that the statistical methods can not be relied on alone for the identification of effective or ineffective schools. Consequently, the study argues that the statistical methods could be employed to identify schools that are unusually effective (or ineffective), as a first step in qualitative research.

Moreover, this study argues that, because a substantial amount of variance is left unexplained within the school, future research on school effectiveness of the primary schools in South Australia should focus on what is happening within the school, at the class-level. Without focusing on what could be happening within the school, the task of identifying that effective (or ineffective) school will continue to be no easier than searching for a needle in a hay stack, only in this case, it may be the wrong hay stack altogether. It is more like trying to judge which boxes among a group of 482 boxes (the number of schools in the study) that are identical from the outside have valuable (or worthless) contents without opening the boxes!

Obviously, the results from this study should be of interest to both proponents and opponents of the BSTP. More importantly, this study provides information that should shape the general direction upon which the debate on measurement of school effects based on students' scores from the BSTP should follow.

13

References

Adams, R. J. and Khoo, S. T. (1993). *QUEST: The Interactive Test Analysis System* [Computer Software]. Hawthorn, Vic: Australian Council for Education Research.

Afrassa, T. M. (2002). *Changes in Mathematics Achievement Over Time in Australia and Ethiopia*. Flinders University Institute of International Education Research Collection Number 4. Adelaide: Shannon Research Press.

Afrassa, T. M. and Keeves, J. P. (1999). *Factors Influencing Years 3 and 5 Students' Basic Skills Test Performance: A Two Level HLM Analysis*. Paper Presented to the Research Expo 1999. March 20, 1999. Adelaide.

Aiken, L. S. and West, S. G. (1996). *Multiple Regression: Testing and Interpreting Interactions.* Newbury Park:Sage Publications.

Ainley, J., Goldman, J. and Reed. R. (1990). *Primary Schooling in Victoria*. ACER Research Monogragph Number 37, Hawthorn, Vic: ACER.

Ainley, J., Graetz, B. Long, M. and Batten, M. (1995). *Social Economic Status and School Education*. Canberra: Australian Government Publishing Service.

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

Allerup, P. (1997). Rasch Measurement Theory. In J. P. Keeves (Ed.), *Educational Research, Methodology, and Measurement: An International Handbook* (2nd ed., pp.836-40). Oxford: Pergamon Press.

Andrich, D. and Masters, G. N. (1988). In J. P. Keeves (Ed.), *Educational Research, Methodology, and Measurement: An International Handbook* (pp.297-303). Oxford: Pergamon Press.

Angoff, W. H. (1982). Summary and Derivation of Equating Methods Used at ETS. In P. W. Holland and D. B. Rubin (Eds.), *Test Equating* (pp.57-69). New York: Academic Press.

Baker, A. P., Xu, Dengke and Detch, E. (1995). *The Measure of Education: A Review of the Tennessee Value Added Assessment System*, Office of Education

Accountability, Tennessee Department of Education, Nashville, TN: Comptroller for the Treasury.

Baker, F. B. and Al-Karni, A. (1993). A Comparison of Two Procedures for Computing IRT Equating Coefficients. *Journal of Educational Measurement, 28* (2), 147-162.

Banerji, M. (2000). Construct Validity of Scores/Measures from a Developmental Assessment in Mathematics Using Classical and Many-Facet Rasch Measurement*. Journal of Applied Measurement, 1* (2), 177-98.

Barnard, J. J. (1999). Item Analysis in Test Construction. In G. N. Masters, and J. P. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 195-206). Oxford: Pergamon.

Bejar, I. I. (1983). *Achievement Testing: Recent Advances*. Beverly Hills, CA: Sage Publications.

Bereiter, C. (1963). Some Persisting Dilemmas in the Measurement of Change. In C. W. Harris (Ed.), *Problems in the Measurement of Change* (pp. 3-20). Madison: University of Wisconsin Press.

Bishop, F. and Clements, M. A. (1994). *Predictions of Gender Differences in Performance of Years 5 and 6 Children on Pencil-and-Paper Mathematics Items.* Paper Presented at the Seventeenth Annual Conference of Mathematics Education Research Group of Australasia held at the South Cross University. 5-8 July 1994. Lismore, Australia.

Blane, D. (1985). A Longitudinal Study of Children's School Mobility and Attainment in Mathematics. *Educational Studies in Mathematics, 16* (2), 127-142.

Blomqvist, N. (1977). On the Relation between Change and the Initial Value. *Journal of the American Statistical Association, 72,* 746-749.

Bloom, B. S. (1976). *Human Characteristics and School Learning*. New York: McGraw-Hill.

Bock, R. D. and Wolfe, R. (1996). *Audit and Review of the Tennessee Value Added Assessment System (TVAAS): Final Report,* Chicago, IL: University of Chicago.

Bogan, E. D, and Yen, W. M. (1983). *Detecting Multidimensionality and Examining Its Effects on Vertical Equating with the Three-Parameter Logistic Model.* Paper Presented at the Annual Meeting of the American Educational Research Association. April 11-15, 1983. Montreal, Quebec.

Bolt, D. M. (1999). Evaluating the Effects of Multidimensionality on IRT True-score Equating. *Applied Measurement in Education, 12* (4), 383-407.

Bond T. G. and Fox, C. M. (2001). *Applying the Rasch Model. Fundamental Measurement in the Human Sciences*. Maharah, NJ: Lawrence Erlbaum and Associates.

Bosker, R. J. and Akkermans, L. M. W. (1994*). Educational Effectiveness Models*. Enschede: University of Twente/OCTO.

Bosker, R. J. and Scheerens, J. (1989). Issues in the Interpretation of the Results of School Effectiveness Research. In B. P. M. Creemers and J. Scheerens (Eds.), *Development in School Effectiveness Research*, Special Issue of *International Journal of Educational Research*, *13* (7), 741-751.

Bosker, R. J. and Witziers, B. (1996). *The Magnitude of School Effects, or: Does it Really Matter Which School a Student Attends?* Paper Presented at the Annual Meeting of the American Educational Research Association. New York, NY.

Bourke, S. (1998). *School Level Variables as Predictors of Individual Student Achievement*. Paper Presented at the Annual Conference of Australian Association for Research in Education. Nov/Dec. 1998. Adelaide.

Bourke, S. P., Mills, J. K., Stanyon, J. and Holzer, F. (1981). *Performance in Literacy and Numeracy:* 1980. Canberra: AGPS.

Braggett, K. E. (1997). *Reading Development and Instructional Practices within High and Low Achievement Primary Schools*. Unpublished Doctoral Thesis, University of Newcastle.

Brandsma, H. and Doolaard, S. (1999). Differences in Effectiveness between Primary Schools and their Impact on Secondary School: Recommendations. *School Effectiveness and School Improvement, 10* (4), 430-450.

Brandsma, H. P. and Knuver, A. V. M. (1989). Effects of School Classroom Characteristics on Pupil Progress in Language and Arithmetic. In B. P. M. Creemers and J. Scheerens (Eds.), *Development in School Effectiveness Research*, Special issue of *International Journal of Educational Research*, *13* (7), 777-788.

Brent, G. and Diobilda, N. (1993). Effects of Curriculum Alignment versus Direct Instruction on Urban Children. *Journal of Education Research*, *86* (6), 333-338.

Bressoux, P. (1995). The Effects of the School Context on Children Learning – The Effects of Class and School on Reading. *Revue Francaise de Sociologie, 36* (2), 273.

Brewer, D. (1998). Dare to Change: Can Teachers Make a Difference? *Education and Society, 16* (2), 79-83.

Browne, W., Healy, M., Cameron, B. and Charlton, C. (2001). *MLwiN Version 1.10.0007, Multilevel Model Project*. London: University of London, Institute of Education.

Bryk, A. S. and Raudenbush, S. W. (1987). Applications of Hierarchical Linear Models to Assessing Change. *Psychological Bulletin*, *101* (1), 147 - 158.

Bryk, A. S. and Raudenbush, S. W. (1992). *Hierarchical Linear Models: Application and Data Analysis Methods*. Newbury Park: Sage Publication.

Bryk, A. S., Raudenbush, S. W. and Congdon, R. T. (1994). *Hierarchical Linear Modeling with the Hierarchical Linear Modeling/2L and Hierarchical Linear Modeling/3L Programs* [Computer Software]. Chicago: Scientific Software International.

Bryk, A. S., Raudenbush, S. W. and Congdon, R. T. (1996). *HLM: Hierarchical Linear Modeling and Non-linear Modeling with HLM/2L and HLM/3L Programs* [Computer Software]. Chicago: Scientific Software International.

Camilli, G. (1993). Scale Shrinkage in Vertical Equating. *Applied Psychological Measurement*, *17* (4), 379-88.

Carroll, J. B. (1963). A Model of School Learning. *Teachers College Record*, 64, 723-33.

Carroll, J. B. and Spearritt, D. (1967). *A Study of a "Model of School Learning."* Monograph Number 4. Cambridge, MA: Harvard University Press.

Cheung, K. C., Keeves, J. P., Sellin, N. and Tsoi, S. C. (1990). The Analysis of Multilevel Data in Educational Research: Studies of Problems and Their Solutions. *International Journal of Educational Research, 14* (3), 215-319.

Choppin, B. H. (1997). Objective Tests. In J. P. Keeves (Ed.), *Educational Research, Methodology, and Measurement: An International Handbook* (2nd ed., pp.771-774). Oxford: Pergamon Press.

Choppin, B. H. and Wolf (1992). Correction for Guessing. In J. P. Keeves (Ed.), *The IEA Technical Handbook*. The Hague: IEA.

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

Cohen, J. (1992). A Power Primer. *Psychological Bulletin, 112* (1), 155-159.

Coleman, J. S., Cambell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D. and York, R. L. (1966). *Equality of Education Opportunity.* Washington: US Government Publishing Office.

Coley, R. J. (2002). *An Even Start: Indicators of Inequality in School Readiness*, Princeton: Education Testing Services. [Online]. Website: http://www.ets.org/research/pic/Unevenstart.pdf [05 December 2002].

Comber, L. C. and Keeves, J. P. (1973). *Science Education in Nineteen Countries: An Empirical Study*. Stockholm: Amquist and Wiksell.

Cook, L. L., and Eignor, D. R. (1991). An NCME Instructional Module on IRT Equating Method. *Educational Measurement: Issues and Practice*, *10* (3), 37-45.

Creemers, B. P. M. (1994). *The Effective Classroom*, London: Redwood Books.

Creemers, B. P. M. and Reezigt, G. J. (1996). School Level Conditions Affecting the Effectiveness of Instruction*. School Effectiveness & School Improvement*, *7* (3), 197-228.

Creemers, B., Scheerens, J. and Reynolds, D. (2001). Theory Development in School Effectiveness Research. In C. Teddlie and D. Reynolds (Eds.), *The International Handbook of School Effectiveness Research* (pp. 283-298). London: Routledge Falmer.

Crone, L. J., Lang., M. and Franklin, B. J. (1994). *Achievement Measures of School Effectiveness: Comparison of Consistency Across Years*. Paper Presented at the Annual Meeting of the American Educational Research Association. 4-8 April. New Orleans.

Cuttance, P. (1985). Framework for Research on the Effects of Schooling. In D. Reynolds (Ed.), *Studying School Effectiveness*, Lewes: Falmer Press.

Cuttance, P. (1987). *Modelling Variation in the Effectiveness of Schooling*. Edinburgh: CES.

Davies, A. (1991). Performance of Children from non-English Speaking Background on the New South Wales Basic Skills Tests of Numeracy: Issues of Test Bias and Language Proficiency. *Culture and Curriculum (UK)*, *4* (2), 149-161.

De Leeuw, J. and Kreft, I. G. G. (1995a). Questioning Multilevel Models. In I. G. G. Kreft (Guest Editor) Hierarchical Linear Models: Problems and Prospects. Special Issue of *Journal of Educational and Behavioral Statistics*, *20* (2), 171-189.

De Leeuw, J. and Kreft, I. G. G. (1995b). Not Much Disagreement, it Seems. In I. G. G. Kreft (Guest Editor) Hierarchical Linear Models: Problems and Prospects. Special Issue of *Journal of Educational and Behavioral Statistics, 20* (2), 239-240.

DeMars, C. (2001). Group Differences Based on IRT Scores: Does the Model Matter? *Educational and Psychological Measurement*, *61* (1), 60-70.

DeMars, C. (2002). Incomplete Data and Item Parameter Estimates under JMLE and MML Estimation. *Applied Measurement in Education*, *15* (1), 15-31.

Dorans, N. J. (1990). Equating Methods and Sampling designs. *Applied Measurement in Education*, *3* (1), 3-17.

Dorans, N. J. and Kingston N. M. (1985). The Effects of Violations of Unidimensionality on the Estimation of Item and Ability Parameters and on Item Response Theory Equating of the GRE Verbal Scale. *Journal of Educational Measurement*, *22* (4), 249-262.

Douglas, G. (1988). Latent Trait Measurement Model. In J. P. Keeves (Ed.), *Educational Research, Methodology, and Measurement: An International Handbook* (pp.282-286). Oxford: Pergamon Press.

Douglas, K. (1988). *The Effects of Specification Error on Regression Based Procedures used in Assessment of School Merit*. Ph.D. Thesis, Florida State University.

Draper, D. (1995). Inference and Hierarchical Modeling in Social Sciences. *Journal of Educational and Behavioral Statistics*, 20 (2), 115-147.

Ehrenberg, R. G., Brewer, A. G., Gamoran, A. and Willms, J. D. (2001). Class Size and Student Achievement. *Journal of the American Psychological Society*, *2* (1), 1-29.

Einsiedler, W. and Treinies, G. (1997). The Effects of Teaching Methods, Class Effects, and Patterns of Cognitive Teacher-Pupil Interactions in an Experimental Study in Primary School Classes. *School Effectiveness & School Improvement, 8* (3), 327-353.

Embretson, S. E. (1995). Implications of a Multidimentional Latent Trait Model for Measuring Change. In L. M. Collins and J. L. Horn (Eds.), *The Best Methods of Measuring Change* (pp. 184-197). Washington DC: American Psychological Association.

Embretson, S. E. (1999). Issues in the Measurement of Cognitive Abilities. In S. E. Embretson and S. L. Hershberger (Eds.), *The New Rules of Measurement: What Every Psychologist and Educator Should Know*. Mahwah, NJ: Lawrence Erlbaum Associates.

Engelhard, G. (1980). *An Introduction to Rasch Measurement and Its Application to Test Equating in the Comprehensive Assessment Program*. Paper presented at the Annual Meeting of the Northern Illinois Association for Educational Research, Evaluation and Development. Bloomingdale, IL.

Ethington, C. A. (1992). Gender Differences in A Psychological Model of Mathematics Achievement. *Journal for Research in Mathematics Education*, *23* (2), 166-181.

Fields, B. A. (1995). Family Mobility: Social and Academic Effects on Young Adolescents. *Youth Studies Australia*, *14* (2), 27-31.

Fields, B. A. (1997a). The Social and Educational Effects of Student Mobility: Implications for Teachers and Guidance Officers. *Australian Journal of Guidance and Counseling*, *7* (1), 45-56.

Fields, B. A. (1997b). Children on the Move: The Social and Educational Effects of Family Mobility: *Children Australian*, *22* (3), 4-9.

Fischer, G. H. and Molenaar, I. W. (1995). *Rasch Models: Foundations, Recent Developments, and Applications*. New York: Springer-Verlag.

Fitz-Gibbon, C. T. (1990). *Performance Indicators: A BERA Dialogue*. Clevedon, Avon: Multi-lingual Matters.

Fitz-Gibbon, C. T. (1991). Multilevel Modelling in an Indicator System. In S. W. Raudenbush and J. D. Willms (Eds.), *Schools, Classrooms, and Pupils: International Studies of School from a Multilevel Perspective* (Chapter 6). San Diego: Academic Press.

Fitz-Gibbon, C. T. (1996). *Monitoring Education: Indicators, Quality and Effectiveness,* London: Cassell.

Fitz-Gibbon, C. T. (1997). *The Value Added National Project*, London: SCAA.

Fitz-Gibbon, C. T. and Kochan, S. (2001). School Effectiveness and Education Indicators. In C. Teddlie and D. Reynolds (Eds.), *The International Handbook of School Effectiveness Research* (pp. 257-282). London: Routledge Falmer,.

Fuchs, L. S., Fuchs, D., Karns, K., Hamlett, C. L., Dutka, S. and Katzaroff, M. (2000). The Importance of Providing Background Information on the Structure and Scoring of Performance Assessments. *Applied Measurement in Education*, *13* (1), 1-34.

Gage, N. J. (Ed.), (1963). *Handbook of Research on Teaching*. Chicago: Rand McNally.

Gialluca, K. A. (1984). *Methods for Equating Mental Tests*. *Interim Report for Period March 1982-October.* St. Paul, MN: Assessment Systems Corp.

Gill, S. and Reynolds, A. J. (1999). Educational Expectation and School Achievement of Urban African American Children. *Journal of School Psychology*, *37* (4), 403-424.

Glowacki, M. L. (1991). *An Analysis of Test Equating Models for the Alabama High School Graduation Examination*. Paper Presented at the Annual Meeting of the Mid-South Educational Research Association. Lexington, KY.

Goh, S. C. and Fraser, B. J. (1996*). Classroom Climate and Student Outcomes in Elementary Mathematics.* Paper Presented at the Joint Conference of Educational Research Association, Singapore and Australian Association for Research in Education. 25-29 November 1996. Singapore.

Goldstein, H. (1987). *Multilevel Models in Educational Social Research*, London: Charles Griffin.

Goldstein, H. (1991). Better Ways of Comparing Schools? *Journal of Educational Statistics*, *16*, 89-91.

Goldstein, H. (1997). Methods in School Effectiveness Research*. School Effectiveness & School Improvement*, *8* (4), 369-395.

Goldstein, H., Rasbash, J., Yang, H., Woodhouse, G., Pan, H., Nuttall, D. and Thomas, S. (1992). Multilevel Models for Comparing Schools. *Multilevel Modeling Newsletter, 4* (2), 5-6.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D. and Thomas, S. (1993). Multilevel Analysis of School Examination Results. *Oxford Review of Education*, *19*, 425-433.

Gray, J., Goldstein, H. and Jesson, D. (1996). Changes and Improvement in Schools' Effectiveness: Trends Over Five Years. *Research Papers in Education,* 11, 35-51.

Gray, J., Hopkins, D., Reynolds, D., Wilcox, B., Farrell, S and Jesson, D. (1999). *Improving Schools: Performance and Potential*, Buckingham: Open University Press.

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 & School Improvement*, *6* (2), 97-114.

Griffin, M. and Harvey, D. (1995). When Do Principals and Teachers Think Children Should Start School? *Australian Journal of Early Childhood*, *20* (3), 27-32.

Gustafsson, J. E. (1979). The Rasch Model in Vertical Equating of Tests: A Critique of Slinde and Linn. *Journal of Educational Measurement*, *16* (3), 153-58.

Hambleton, R. K. (1989). Principles and Selected Applications of Item Response Theory. *Education Measurement*. (3rd ed.), New York: Macmillan.

Hambleton, R. K. and Swaminathan, H. (1985). *Item Response Theory: Principles and Application*. Boston, MA: Kluwer Academic.

Hambleton, R. K., Swaminathan, H., Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park: Sage Publications.

Harker, R. and Nash, R. (1996). Academic Outcomes and School Effectiveness — Type A and Type B Effects. *New Zealand Journal of Educational Studies, 31* (2), 143-170.

Harnischfeger, A. and Wiley, D. T. (1976). The Teaching and Learning Process in Elementary Schools: A Synoptic View. *Curriculum Inquiry*, *6*, 5-43.

Harnischfeger, A. and Wiley, D. T. (1978). Conceptual Issues in Models of School. *Curriculum Studies*, *10* (30), 215-231.

Hattie, J. (1985). Methodology review: Assessing Unidimensionality of Tests and Items. *Applied Psychological Measurement*, *9* (2), 139 - 164.

Hattie, J. (1992). Measuring the Effects of Schooling. *Australian Journal of Education*, *36* (1), 5-13.

Heyl, E. (1996). *Teacher's Networks, Structure and Influence of Collegial Contact within Schools.* Enschede: University of Twente.

Hill, P. W. (1996). *The Value Schools Add to the Learning Achievement of their Students*. Paper Presented at the One-day Invitational Conference 'School of the Third Millennium'. February 1996.

Hill, P. W. and Goldstein, H. (1998). Multilevel Modeling of Educational Data with Cross-Classification and Missing Identification for Units. *Journal of Educational and Behavioral Statistics*, *23* (2), 117-128.

Hill, P. W. and Rowe, K. J. (1998). Modelling Student Progress in Studies of Educational Effectiveness. *School Effectiveness & School Improvement*, *9* (3), 310-333.

Hill, P. W., Rowe, K. J. (1996). Multilevel Modelling in School Effectiveness Research. *School Effectiveness and School Improvement, 7* (1), 1-34.

Hill, P. W., Rowe, K. J., Holmes, S. P. (1996). *Modelling Student Progress*. Paper Presented at the International Congress for School Effectiveness and Improvement. January 1996. Minsk, Republic of Belarus.

Hillman, K., Marks, G. and McKenzie, P. (2002). *LSAY Briefing: Rural and Urban Differences in Australian Education*. Melbourne: ACER.

Holmes, S. E. (1982). Unidimensionality and Vertical Equating with the Rasch Model. *Journal of Educational Measurement*, *19* (2), 139-47.

Howie, S. (2002). *English Language Proficiency and Contextual Factors Influencing Mathematics Achievement of Secondary School Pupils in South Africa*. Enschede: Print Partners Ipskamp.

Hox, J. J. (1995). *Applied Multilevel Analysis*. Amsterdam: TT-Publikaties. [Online]. Website: http://www.fss.uu.nl/ms/jh/publist/amaboek.pdf [21 December, 2002].

Hungi, N. (1997). *Measuring Basic Skills across Primary School Years.* Unpublished Master of Education Thesis. Adelaide: Flinders University of South Australia.

Hungi, N. (2003). *Measuring the Value-added by Schools to Student Achievement across Primary School Grades*. Unpublished Ph.D. Thesis. Adelaide: Flinders University of South Australia

Husén, T. (Ed.), (1967). *International Study of Achievement in Mathematics* (Vol.2). Stockholm: Almqvist and Wiksell.

Jaeger R. M. (1981). Some Exploratory Indices for Selection of a Test Equating Method. *Journal of Educational Measurement*, *18* (1), 23-38.

Jaeger, R. M. (1980). *Some Exploratory Indices for the Selection of a Test Equating Method*. Paper Presented at the Annual Meeting of the American Education Research Association. Boston, MA.

Jencks, C, Smith, M., Acland, H., Bane, M. J., Cohen, D., Gintis, H., Heyns, B. and Michelson, S. (1972). *Inequality: A Reassessment of the Effect of Family and Schooling in America*. New York: Basic Books.

Jesson, D. and Gray, J. (1991). Slants on Slopes: Using Multi-level Models to Investigate Differential School Effectiveness and its Impact on Pupils' Examination Results. *School Effectiveness and School Improvement*, *2* (3), 230-247.

Jolly, D. V. and Deloney, P. (1993). *Alternative Organizational Plans: Options for Consideration.* Austin, TX: Southwest Educational Development Lab.

Keats, J. A. (1997). Classical Test Theory. In J. P. Keeves (Ed.), *Educational Research, Methodology, and Measurement: An International Handbook* (2nd ed., pp.713-19). Oxford: Pergamon Press.

Keeves, J. P. and Alagumalai, S. (1999). New Approaches to Measurement. In G. N. Masters and J. P. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 23-42). Oxford: Pergamon.

Keeves, J. P. (1975). The Home, the School and Achievement in Mathematics and Science. *Science Education*, *59* (4), 207-218.

Keeves, J. P. (1986). *Equitable Opportunities in Australian Education*. Melbourne: Ministry of Education (School Division), Victoria.

Keeves, J. P. (1992a). Scaling Achievement Test Scores. *The IEA Technical Handbook* (pp.107-125). The Hague: IEA.

Keeves, J. P. (1992b). The Design and Conduct of the Second Science Study. In J. P. Keeves (Ed.), *The IEA Study of Science III: Changes in Science Education and Achievement: 1970 to 1984* (pp. 42-67). Oxford: Pergamon Press.

Keeves, J. P. (1995). *The World of Learning: Selected Key Findings from 35 Years of IEA Research*. The Hague: The International Association for the Evaluation of Educational Achievement (IEA).

Keeves, J. P. (1997). Suppressor Variables. In J. P. Keeves (Ed.), *Educational Research, Methodology, and Measurement: an International Handbook* (pp.695-6). Oxford: Pergamon Press.

Keeves, J. P. and Bourke, S. F. (1976). *Literacy and Numeracy in Australian Schools: A First Report. Australian Studies in School Performance Volume I.* Canberra: Australian Government Publishing Service.

Keeves, J. P. and Saha, L. J. (1992). Home Background Factors and Educational Outcomes, In J. P. Keeves (Ed.), *The IEA Studies of Science III. Changes in Science Education and Achievement. 1970-1984* (pp. 165-186). Oxford: Pergamon.

Keeves, J. P. and Schleicher, A. (1992). Changes in Science Achievement: 1970-84. In J. P. Keeves (Ed.), *The IEA Study of Science III: Changes in Science Education and Achievement: 1970 to 1984* (pp. 165 – 186). Oxford: Pergamon.

Keeves, J. P., Matthews, J. K. and Bourke, S. F. (1978). *Educating for Literacy and Numeracy in Australian Schools*. Hawthorn, Vic: Australian Council for Education Research.

Keller, I. M. (1983). Motivational Design of Instruction. In C. Reigeluth (Ed.), *Instructional Design Models*. Hillsdale, NJ: Erlbaum.

Kennedy, E. and Mandeville, G. (2001). Some Methodological Issues in School Effectiveness Research. In C. Teddlie and D. Reynolds (Eds.), *The International Handbook of School Effectiveness Research* (pp. 189-205). London: Routledge Falmer,.

Kennedy, E., Stringfield, S. and Teddlie, C. (1993). Schools do Make a Difference, In C. Teddlie and S. Sprinfield (Eds.), *Schools Make a Difference: A Lesson Learned from a Ten-year Study of School Effects*. New York: Teachers College Press.

Kenyon, D. M. and Stansfield, C. W. (1992). *Extending a Scale of Language Proficiency Using Concurrent Calibration and the Rasch Model.* Paper Presented at the Annual Meeting of the American Educational Research Association (73rd, San Francisco, CA, April 20-24, 1992). Center for Applied Linguistics, Washington, DC.

Kings, R. J. R. (1985). *Children in Double Jeopardy: The Congruence of Family Mobility and School Mobility*. In 'Educational Research: Then and now'

Collected Papers of the Annual Conference, November 1985, (pp.127-130), Hobart: Australian Association for Research in Education.

Kline, P. (1993). Rasch Scaling and other Scales. *The Handbook of Psychological Testing* (pp.68-71). London: Routledge.

Kolen, M. J. (1994). Equating of Tests. In T. Husén and T. N. Postlethwaite (Eds.), *The International Encyclopedia of Education*. (2nd ed., pp. 6349-6355). New York: Pergamon.

Kolen, M. J. (1997). Equating of Tests. In J. P. Keeves (Ed.), *Educational Research, Methodology, and Measurement: An International Handbook* (2nd ed., pp.730-37). Oxford: Pergamon Press.

Kotte, D. (1992). Gender Differences in Science Achievement in 10 Countries. Frankfurt am Main: Peterlang.

Kreft, I. G. G. (1995). *The Effects of Centering in Multilevel Analysis: Is the Public School the Loser or the Winner? A New Analysis of an Old Question*. Paper Presented at AERA 1995, Section D. April 18-22 1995. San Francisco.

Kreft, I. G. G. (1996). *Are Multilevel Techniques Necessary? An Overview, Including Simulation Studies.* CSU, June 25, 1996.

Kreft, I. G. G., De Leeuw, J. and Aiken, L. S. (1995). The Effects of Different Forms of Centering in Hierarchical Linear Models. *Multivariate Behavioral Research*, *30* (1), 1-21.

Kreft, I. G. G., De Leeuw, J. and Kim. K. S. (1990). *Comparing Four Different Statistical Packages for Hierarchical Linear Regression: GENMOD, MM, ML3 and VARCL* (CSE Tech. Rep. 311), Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Kreft, I. G. G., De Leeuw, J. and van der Leeden, R. (1994). Review of Five Multilevel Analysis Programs: BMDP-5V, GENMOD, MM, ML3, VARCL. *The American Statistician, 48*, 324-335.

Kyriakides, L., Gagatsis, A. and Campbell, R. J. (2000). The Significance of the Classroom Effect in Primary Schools: An Application of Creemers' Comprehensive Model of Educational Effectiveness. *School Effectiveness and School Improvement, 11* (4), 501-529.

Lake, D. (1998). Assessment of the Suitability of a Western Instrument in a Non-western Environment using Rasch Analysis*. Education Research and Perspectives*, *25* (2), 68-82.

Lamdin, D. J. (1995). Testing for the Effect of School Size on Student Achievement within a School District. *Education Economics*, *3* (1), 33-42.

Lawley, D. N. (1942). *On Problems Connecting with Item Selection and Test Construction* (pp.273-287). Edinburgh, UK: Moray House, University of Edinburgh.

Lietz, P. (1996). *Changes in Reading Comprehension across Culture and Over Time*. Minister: Waxman.

Linacre, J. M. and Wright, B. D. (1989). *The "Length" of a Logit. Rasch Measurement Transactions*, 3(2) 54-5. [Online]. Website: http://www.rasch.org/ rmt/rmt32.htm [15 January, 2001].

Lokan, J. and Greenwood, L. (2001). *Maths and Science on the Line. Australian Year 12 Students' Performance in the Third International Mathematics and Science Study*. Melbourne: Australian Council for Educational Research.

Lokan, J., Ford, P. and Greenwood, L. (1996). *Maths and Science on the Line. Australian Junior Secondary Students' Performance in the Third International Mathematics and Science Study*. Melbourne: Australian Council for Educational Research.

Lokan, J., Ford, P. and Greenwood, L. (1997). *Maths and Science on the Line. Australian Middle Primary Students' Performance in the Third International Mathematics and Science Study*. Melbourne: Australian Council for Educational Research.

Lokan, J., Greenwood, L. and Cresswell, J. (2001). *15-up and Counting, Reading, Writing, Reasoning … How Literate are Australia's Students? PISA 2000 Survey of Students' Reading, Mathematical and Scientific Literacy Skills*, Melbourne: Australian Council for Educational Research.

Longford, N. T. (1990). *VARCL Software for Variance Component Analysis of Data with Nested Random Effects (Maximum Likelihood)* [Computer Software]. Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). *Application of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M. (1982). Item Response Theory and Equating–Technical Summary. In P. W. Holland and D. B. Rubin (Eds.), *Test Equating* (pp. 141-148). Princeton, NJ: Educational Testing Service, Academic Press.

Lord, F. M. and Stocking, M. L. (1988). Item Response Theory. In J. P. Keeves (Ed.), *Educational Research, Methodology, and Measurement: an International Handbook* (pp.269-272). Oxford: Pergamon Press.

Loyd, B. H. and Hoover, H. D. (1980). Vertical Equating Using the Rasch Model. *Journal of Educational Measurement*, *17* (3), 179-93.

Luyten, H. (1994a). *Stability of School Effects in Secondary Education: The Impact of Variance across Subjects and Years*. Paper Presented at the Annual Meeting of the American Educational Research Association. 4-8 April. New Orleans.

Luyten, H. (1994b). *School Size Effects on the Achievement in Secondary Education. Evidence from the Netherlands, Sweden and the USA*. Paper Presented at the Annual Meeting of the American Educational Research Association. April 4-8, 1994. New Orleans, LA.

Luyten, H. and De Jong, R. (1998). Parallel Classes: Differences and Similarities: Teacher Effects and School Effects in Secondary Schools. *School Effectiveness and School Improvement*, *9* (4), 437-473.

Mandeville, G. K. (1988). School Effectiveness Indices Revisited: Cross-year Stability. *Journal of Educational Measurement*, *25*, 349-365.

Mandeville, G. K. and Anderson, L. W. (1987). A Study of the Stability of School Effectiveness Measures across Grade Level and Subjects. *Journal of Educational Measurement*, *24*, 203-216.

Marks, G. and Ainley, J. (1997). *Reading Comprehension and Numeracy among Junior Secondary Schools Students in Australia*, LSAY Research Report Number 3, Melbourne: Australian Council for Educational Research.

Marks, G., Fleming, N., Long, M. and McMillan, J. (2000). *Patterns of Participation in Year 12 and Higher Education in Australia: Trends and Issues,* LSAY Research Report No. 17, Melbourne: ACER. [Online]. Website: http://www.acer.edu.au/research/vocational/lsay/reports/lsay17.pdf [18 December, 2002].

Marsh, H. W. (1998). *Two Multilevel Modeling to Evaluate Change Over Time: Regression to the Mean Biases*. Sydney: UWS.

Marsh, H. W. and Hocevar, D. (1983). Application of Confirmatory Factor Analysis to the Study of Self-concept: First- and High-order Factor models and their Invariance Across Groups. *Psychological Bulletin*, *97* (3), 562-582.

Marsh, H. W. and Yeung, A. S. (1999). Longitudinal Structural Equation Models of Academic Self-concept and Achievement: Gender Differences in the Development of Math and English Constructs. *American Educational Research Journal*, *35* (4), 705-738.

Martin, J. and Meade, P. (1979). *The Educational Experience of Sydney High School Students: Report Number 1*. Canberra: Australian Government Publishing Services.

Masters, G. N. (1994). *National Comparable Achievement Measures: Principles and Options*. Hawthorn, Vic: ACER (mimeo).

Masters, G. N. and Forster, M. (1997). *Mapping Literacy Achievement. Results of the 1996 National School Literacy Survey*, Canberra: Department of Education, Training and Youth Affairs.

Masters, G. N. and Foster, M. (2000). *The Assessments We Need. Australian Council for Educational Research*. [Online]. Website: http://www.acer.edu.au/ [26 August, 2002].

Masters, G. N. and Keeves, J. P. (1999). *Advances in Measurement in Educational and Psychological Research and Assessment.* Oxford: Pergamon.

McKenzie, P. (1988). Secondary School Size, Operating Cost and Curriculum Structure. In *Educational Research in Australia: Indigenous or Exotic*? Papers Presented at the Annual Conference of the Australian Association for Research in Education held at the University of New England. 30 Nov. – 4 Dec. 1988. Armidale, New South Wales.

McNamara, T. F. (1996). *Measuring Second Language Performance*. New York: Addison Wesley Longman.

McPherson, A. (1993). *Measuring Added Value in Schools* (pp. 1-4). UK: National Commission on Education.

Meyer, R. H. (1996). Value-Added Indicators of School Performance. In A. E. Hanushek and W. D. Jorgenson (Eds.), *Improving America's Schools: The Role of Incentives* (pp. 197-223). Washington, DC: National Academic Press.

Meyer, R. H. (1997). Value-Added Indicators of School Performance - A Primer. *Economics of Education Review, 16* (3), 283-301.

Mills, G. (1986). Changing Schools Again. *Education News*, *19* (9), 12-17.

Mohandas, R. (1996). *Test Equating, Problems and Solutions.* Unpublished MEd Thesis, School of Education. Adelaide: The Flinders University of South Australia.

Mohandas, R. (1999). *Mathematics and Science Achievement of Junior Secondary School Students in Indonesia*. Unpublished Ph.D. Thesis. The Flinders University of South Australia.

Mok, M. F. and Flynn, M. (1996). School Size and Academic Achievement in the HSC Examination: Is there a relationship? *Issues In Educational Research*, 6(1), 57-78. [Online]. Website: http://education.curtin.edu.au/iier/iier6/mok.html (20 December, 2002].

Morris, C. N. (1995). Hierarchical Models for Educational Data: An Overview. *Journal of Educational and Behavioural Statistics*, *20* (2), 190-200.

Morrison, C. A. and Fitzpatrick, S. J. (1992). *Direct and Indirect Equating*: *A Comparison of Four Methods using the Rasch Model.* Austin, TX: Measurement and Evaluation Center, The University of Texas.

Mortimore, P., Sammons., Stoll, L., Lewis, D. and Ecob, R. (1988*). School Matters*. Wells: Open Books.

Muller, P. A., Stage, F. K. and Kinzie, J. (2001) Science Achievement Growth and Trajectories: Understanding Factors Related to Gender and Racial-ethnic Differences in Pre-college Science Achievement. *American Educational Research Journal*, *38* (4), 981-1012.

Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. L. and Smith, T. A. (1997). *Mathematics Achievement in Primary School Years. IEA'S Third International Mathematics and Science Study*, Chestnut Hill, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Gregory, K. D., Garden, R. A., O'Connor, K. M., Chrostowski, S. J. and Smith, T. A. (2000). *TIMSS 1999 International Mathematics Report. Findings from the IEA'S Repeat of the Third International Mathematics and Science Study at the Eighth Grade*, Chestnut Hill, MA: Boston College.

Nuttall, D. (1990). *Differences in Examination Performance*, RS 1277/90, London: Research and Statistics Branch, ILEA.

Nuttall, D., Goldstein, H., Prosser, R. and Rasbash, J. (1989). Differential School Effectiveness. *International Journal of Educational Research, 13* (7), 769-776.

O'Brien, M. L. and Tohn, D. (1984). *Applying and Evaluating Rasch Vertical Equating Procedures for Out-of-Level Testing*. Paper Presented at the Annual Meeting of the Eastern Educational Research Association. February 10, 1984. West Palm Beach, FL.

Organisation for Economic Co-operation and Development (2001). *Knowledge and Skills for Life. First Results from PISA 2000*. Paris: OECD.

Osterlind, S. J. (1983). Test Item Bias. *Sage University Paper Series On Quantitative Application in The Social Sciences*, *07-001*. Beverly Hills: Sage Publications.

Paterson, L. (1991). Trends in Achievement in Scottish Secondary Schools. In S. W. Raudenbush, and J. D. Willms (Eds.), *Schools, Classrooms, and Pupils: International Studies of School from a Multilevel Perspective* (pp. 85-100). San Diego: Academic Press.

Peaker, G. F. (1967). The Regression Analyses of the National Survey. In Central Advisory Council for Education 1967, *Children and their Primary Schools. A Report of the Central Advisory Council for Education. Vol. 2: Research and Theories*. London: HMSO.

Peaker, G. F. (1975). *An Empirical Study of Education in Twenty-One Countries: A Technical Report*. Stockholm: Almqvist & Wilsell.

Peck, R. G. and Trimmer, K. J. (1995). The Late Birthday Effect in Western Australia. *Issues in Education Research*, *5* (1), 35-52.

Petersen, N. S., Kolen, M. J. and Hoover, H. D. (1989). Scaling, Norming, and Equating. In R. L. Linn (Ed.), 1989 *Educational Measurement*. (3rd ed., pp.221-262). New York: Macmillan.

Phillips, S. E. (1986). The Effects of the Deletion of Misfitting Persons on Vertical Equating via the Rasch Model. *Journal of Educational Measurement*, *23* (2), 107-18.

Pituch, K. A. (1999). Describing School Effects with Residuals Terms. *Evaluation Review, 23* (2), 190-211.

Plecki, M. (1991). *The Relationship between Elementary School Size and Student Achievement*. Paper Presented at the Annual Meeting of the American Educational Research Association. April 3-7, 1991. Chicago, IL.

Porter, R. (1980). The Effects of Preschool Experiences and Family Environment on Children's Cognitive and Social Development at School Entry. *Australian Journal of Early Childhood*, 7(2), 20-23.

Postlethwaite, T. N. and Ross, K. N. (1992). *Effective Schools in Reading. Implications for Educational Planners.* The Hague: IEA.

Postlethwaite, T. N. and Wiley, D. E. (1991). *The IEA Study of Science II. Science Achivement in Twenty-Three Countries*. Oxford: Pergamon Press.

Preece, P. (1989). The Pitfalls in Research on School and Teacher Effectiveness. *Research Papers in Education*, *4* (3), 47-69.

Rahmani, Z. (1985). Smoothing Out the Turbulence. *Education News*, *19* (2), 39-41.

Ramilez, A. (1990). High School Size and Equality of Educational Opportunity. *Journal of Rural and Small Schools*, *4* (2), 12-19.

Rasbash, J., Browne, W., Goldstein., H., Yang, M., et al. (2000). *A User's Guide to MLwiN* (2nd ed.). London: Institute of Education.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. (reprinted 1980), Chicago: University of Chicago Press.

Raudenbush, S. W. and Bryk A. S. (1997). Hierarchical Linear Models. In J. P. Keeves (Ed.), *Educational Research, Methodology, and Measurement: An International Handbook* (pp. 2590-2596). Oxford: Pergamon Press.

Raudenbush, S. W. (1988). Educational Applications of Hierarchical Linear Models: A Review. *Journal of Educational Statistics*, *13* (2), 85-116.

Raudenbush, S. W. (1989). The Analysis of Longitudinal, Multilevel Data. *International Journal of Educational Research*, *13*, 721-740.

Raudenbush, S. W. (1995). Statistical Models for Studying the Effects of Social Context on Individual Development. In J. Gottman (Ed.), *The Analysis of Change* (pp. 165-201). Hillsdale, NJ: Lawrence Erlbaum.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). California: Sage Publications.

Raudenbush, S. W. and Willms, J. D. (1991). The Organisation of Schooling and Its Methodological Implications. In S. W. Raudenbush and J. D. Willms (Eds.), *Schools, Classrooms, and Pupils: International Studies of School from a Multilevel Perspective* (pp. 1-12). San Diego: Academic Press.

Raudenbush, S. W. and Willms, J. D. (1995). The Estimation of School Effects. *Journal of Educational and Behavioral Statistics*, *20* (4), 307-335.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F. and Congdon, R.T. (2000). *HLM5: Hierarchical linear and Nonlinear Modeling* [Computer Software]. Lincolnwood, IL: Scientific Software International.

Raywid, M. A. (1997). Small Schools: A Reform that Works. *Educational Leadership*, *55* (4), 34-39

Reckase, M. D. (1981). *To Use or Not to Use--(The One- or Three-Parameter Logistic Model) That Is the Question*. Paper Presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981). Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office. Missouri Univ., Columbia

Reynolds, A. J. and Walberg, H. J. (1991). A Structural Model of Science Achievement. *Journal of Educational Psychology*, *83* (1), 97-107.

Reynolds, A. J. and Walberg, H. J. (1992). A Process Model of Mathematics Achievement and Attitude. *Journal for Research in Mathematics Education, 23* (4), 306-328.

Reynolds, A. J. and Wolfe, B. (1999). Special Education and School Achievement: An Exploratory Analysis with a Central-city Sample. *Education Evaluation & Policy Analysis*, *21* (3), 249-269.

Reynolds, D. and Teddlie, C. (2001). Reflections on the Critics, and Beyond Them. *School Effectiveness and School Improvements, 12* (1), 99-114.

Roberts, P. and Fawcett, G. (1998). *At Risk: A Socioeconomic Analysis of Health and Literacy among Seniors*. Ottawa: Statistic Canada. [Online]. Website: http://www.nald.ca/nls/ials/atrisk/atrisk.pdf [18 December, 2002].

Rogers, H. J. (1997). Multiple Choice Tests, Guessing in. In J. P. Keeves (Ed.), *Educational Research, Methodology, and Measurement: An International Handbook* (2nd ed., pp.766-71). Oxford: Pergamon Press.

Rogosa, D. (1995). Myths and Methods: "Myths about Longitudinal Research" plus Supplemental Questions. In J. Gottman (Ed.), *The Analysis of Change* (pp. 3-66). Hillsdale, NJ: Lawrence Erlbaum.

Rogosa, D. and Saner, D. (1995). Longitudinal Data Analysis Examples with Random Coefficient Models. *Journal of Educational and Behavioral Statistics*, *20* (2), 149-170.

Rothman, S. (1998). *Factors Influencing Assigned Student Achievement Levels*. Paper Presented at the Annual Conference of the Australian Association for Research in Education. 29 November - 3 December 1998. Adelaide.

Rothman, S. (2000). *Factors Influencing Student Non-attendance: A Multilevel Analysis.* Paper Presented at the Meeting of South Australia Institute for Educational Research. March 9, 2000. Adelaide.

Rothman, S. (2001). School Absence and Student Background Factors: A Multilevel Analysis. *International Education Journal*, *2* (1), 59-68.

Rothman, S. (2002). Student Absence in South Australian Schools. *Australian Educational Researcher*, *29* (1), 69-91.

Rothman, S. (2003). *Achievement in Literacy and Numeracy by Australian 14-Year-Olds, 1975-1998*. LSAY Research Report Number 29. Melbourne: ACER. [Online]. Website: http://www.acer.edu.au/research/vocational/lsay/reports/LSAY29.pdf [06 February, 2003].

Rothman, S., and McMillan, J. (forthcoming). *Influences on Achievement in Literacy and Numeracy*. LSAY Research Report. Melbourne: ACER.

Rowe, K. J., Hill, P. W., Holmes-Smith, P. (1996). Methodological Issues in Educational Performance and School Effectiveness Research: A Discussion with Worked Examples. *Australian Journal of Education*, *39* (3), 217-248.

Rumberger, R. W. and Larson, K. A. (1998). Student Mobility and the Increased Risk of High School Dropout. *American Journal of Education*, *107* (1), 1-35.

Rutter, M. (1983). School Effects on Pupil Progress – Finding and Policy Implications. *Child Development*, *54* (1), 1-29.

Sammons, P., Mortimore, P. and Thomas, S. (1996). Do Schools Perform Consistently across Outcome Areas? In J. Gray, D. Reynolds, C. Fitz-Gibbon and D. Jesson, (Eds.), *Merging Traditions: The Future of Research on School Effectiveness and School Improvement* (pp. 3-29). London: Cassell.

Sammons, P., Nuttall, D. and Cuttance, P. (1993). Differential School Effectiveness: Results from a Re-analysis of the Inner London Educational Authority's Junior School Project Data. *British Educational Research Journal, 19* (4), 381-405.

Saunders, L. (1998). '*Value Added' Measurement of School Effectiveness: An Overview*. Slough: The National Foundation for Educational Research.

Saunders, L. (1999). '*Value Added' Measurement of School Effectiveness: A Critical Review.* Slough: The National Foundation for Educational Research.

Scheerens, J. (1992). *Effective Schooling: Research, Theory and Practice*, London: Cassell.

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

Scheuneman, J. (1979). A method of Assessing Bias in Test Items. *Journal of Educational Measurement*, *16* (3), 143-152.

Schratz, M. K. (1984). *Vertical Equating: An Empirical Study of the Consistency of Thurstone and Rasch Model Approaches*. Paper Presented at the Annual Meeting of the National Council on Measurement in Education. April 24-26, 1984. New Orleans, LA.

Shen, L. (1993). *Constructing a Measure for Longitudinal Medical Achievement Studies by the Rasch Model One-Step Equating*. Paper Presented at the Annual Meeting of the American Educational Research Association. April 12-16, 1993. Atlanta, GA.

Silins, H. and Murray-Harvey, R. (1998). *Student as a Central Concern: The Relationship between School, Students and Outcomes*. Paper Presented at the Annual Conference of the Australian Association for Research in Education. November, 1998. Adelaide.

Sime, N. and Gray, J. (1991). *Struggling for Improvement: Some Estimates of the Contribution of School Effects Over Time*. Paper Presented at the Annual Conference of the British Education Research Association.

Skaggs, G. and Lissitz, R. L. (1986). IRT Test Equating: Relevant Issues and a Review of Recent Research. *Review of Educational Research*, *56*, 495 - 529.

Skaggs, G. L. and Robert, W. (1988). Effect of Examinee Ability on Test Equating Invariance. *Applied Psychological Measurement*, *12* (1), 69-82.

Slee, R., Weiner, G. and Tomlinson, S. (Eds.), (1998). *School Effectiveness for Whom? Challenges to School Effectiveness and School Improvement Movements*. London: Falmer Press.

Slinde, J. A. and Linn, R. L. (1977). Vertically Equated Tests: Fact or Phantom? *Journal of Educational Measurement*, *14* (1), 23-32.

Slinde, J. A. and Linn, R. L. (1978). An Exploration of the Adequacy of the Rasch Model For the Problem of Vertical Equating. *Journal of Educational Measurement*, *15* (1), 23-35.

Slinde, J. A. and Linn, R. L. (1979). A Note on Vertical Equating via the Rasch Model for Groups of Quite Different Ability and Tests of Quite Different Difficulty. *Journal of Educational Measurement*, *16* (3), 159-65.

Smith, R. M. and Kramer, G. A. (1992). A Comparison of Two Methods of Test Equating in the Rasch Model. *Educational and Psychological Measurement*, *52* (4), 835-46.

Sontag, M. L. (1984). Vertical Equating Methods: A Comparative Study of their Efficacy. *DAI*, 45-03B, p.1000.

Spearritt, D. (1997). Factor Analysis. In J. P. Keeves (Ed.), *Educational Research, Methodology, and Measurement: An International Handbook* (2nd ed., pp.528-39). Oxford: Pergamon Press.

Stocking, M. L. (1997). Item Response Theory. In J. P. Keeves (Ed.), *Educational Research, Methodology, and Measurement: An International Handbook* (2nd ed., pp.836-40). Oxford: Pergamon Press.

Stocking, M. L. (1999). Item Response Theory. In G. N. Masters and J. P. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 55-63). Oxford: Pergamon.

Stringfield, S. C. and Slavin, R. E. (1992). A Hierarchical Longitudinal Model for Elementary School Effects. In B. P. M. Creemers and G. J. Reezigt (Eds.), *Evaluation of Educational Effectiveness* (pp.35-69). Groningen: ICO.

Tabachnick, B. G. and Fidell, L. S. (1989). *Using Multivariate Statistics*, (2nd ed.). New York: Harper Collins.

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

Teddlie, C. and Reynolds, D. (Eds.), (2001a). *The International Handbook of School Effectiveness Research*, London: Routledge Falmer.

Teddlie, C. and Stringfield, S. (1993). *Schools Make a Difference: A Lesson Learned from a Ten-year Study of School Effects*. New York: Teachers College Press.

Teddlie, C., Reynolds, D. and Sammons, P. (2001). The Methodology and Scientific Properties of School Effectiveness Research. In C. Teddlie and D. Reynolds (Eds.), *The International Handbook of School Effectiveness Research* (pp. 55-133). London: Routledge Falmer.

Temple, J. A. and Reynolds, A. J. (1999). School Mobility and Achievement: Longitudinal Findings from an Urban Cohort. *Journal of School Psychology*, *37* (4), 355-377.

Thanassoulis, E. (1996). Altering the Bias in Differential School Effectiveness Using Data Envelopment Analysis. *Journal of the Operational Research Society*, *47* (7), 882-894.

Thomas, S. and Mortimore, P. (1996). Comparison of Value Added Models for Secondary School Effectiveness. *Research Papers in Education*, *11*, 5-33.

Thomas, S., Sammons, P., Mortimore, P. and Smees, R. (1997). Stability and Consistency in Secondary Schools' Effects' GCSE Outcomes over Three Years. *School Effectiveness and School Improvement*, *8* (2), 169-197.

Thorndike, R. L. and Thorndike, R. M. (1994). Reliability in Educational and Psychological Measurement. In T. Husén and T. N. Postlethwaite (Eds.), *The International Encyclopedia of Education*, (2nd ed., pp. 4981-4995). Oxford: Pergamon.

Thorndike, R. L. (1973b). *Reading Comprehension in Fifteen Countries*. Stockholm: Almqvist and Wiksell.

Thorndike, R. L. (1973a). Reading as Reasoning. *Reading Research Quarterly, 9*, (2), 135-47.

Thrupp, M. (1999). *School Making a Difference: Lets be Realistic*. Buckingham, UK: Open University Press.

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

Trower, P. and Vincent, L. (1995). *The Value Added National Project Technical Report: Secondary*, London: School Curriculum and Assessment Authority.

Tymms, P. (1994). Monitoring School Performance: A Guide for Educators. *School Effectiveness and School Improvement*, *5* (4), 394-397.

Vijver, F. R. and Poortinga, Y. H. (1991). Testing Across Cultures. In R. K Hambleton and J. N. Zaal, (Eds.), *Advances in Education and Psychological Testing* (pp.277-308). Boston, MA: Kluwer Academic.

Walberg, H. J. (1991). Improving School Science in Advanced and Developing Countries. *Review of Educational Research,* 61(1), 25-69.

Waugh, R. F. (2001). Measuring Ideal and Real Self-Concept on the Same Scale, Based on a Multifaceted, Hierarchical Model of Self-Concept. *Educational and Psychological Measurement*, *61* (1), 85-101.

Webster, B. J. and 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.

Webster, W. J., Mendro, R. L., Bembry, K. L. and Orsak, T. H. (1995). *Alternative Methodology for Identifying Effective Schools*. Paper Presented in a

Distinguished Paper Session at the American Educational Research Association Meeting., April 17-21, 1995. San Francisco, California.

Weiss, D. J. and Yoes, M. E. (1991). Item Response Theory. In R. K. Hambleton and J. N. Zaal, (Eds.), *Advances in Education and Psychological Testing* (pp.69-96). Boston, MA: Kluwer Academic.

Wilcox, R. R. (1988). True-score Models. In J. P. Keeves (Ed.), *Educational Research, Methodology, and Measurement: An International Handbook* (pp.267-269). Oxford: Pergamon Press.

Willett, J. B. (1988). Questions and Answers in the Measurement of Change. In E. Rothkopf (Ed.), *Reviews of Research in Education* (1988-89) (pp. 345-422). Washington, DC: American Educational Research Association.

William, D. S., Shu, J. Y. and Talsima, R. (2000). School Effects Indices: Stability of One- and Two-level Formulations. *The Journal of Experimental Education*, *68* (3), 239-246.

Willms, D. J. and Raudenbush, S. W. (1989). A Longitudinal Hierarchical Linear Model for Estimating School Effects and their Stability. *Journal of Educational Measurement*, *26*, 209-232.

Willms, D. J. and Somers, M. A. (2001). Family, Classroom, and School Effects on Children's Educational Outcomes in Latin America. *School Effectiveness and School Improvement, 12* (4), 409-445.

Willms, J. and Chen, M. (1989). The Effects of Ability Grouping on the Ethnic Achievement Gap in Israeli Elementary Schools. *American Journal of Education, 97* (3), 237-57.

Willms, J. D. (1992). *Monitoring School Performance: A Guide for Educators*. Washington DC: Falmer Press.

Willms, J. D. and Kerckhoff, A. C. (1995). The Challenge of Developing New Educational Indicators. *Educational Evaluation and Policy Analysis*, *17* (1), 113-131.

Witte, J. F. and Walsh, D. J. (1990). A Systematic Test of Effective School Model. *Educational Evaluation and Policy Analysis*, *12*, 188-212.

Woldbeck, T. (1998). *Basic Concepts in Modern Methods of Test Equating*. Paper Presented at the Annual Meeting of the Southwest Psychological Association. April 1998. New Orleans, LA.

Wright, B. D. (1988). Rasch Measurement Models. In J. P. Keeves (Ed.), *Educational Research, Methodology, and Measurement: An International Handbook* (pp.286-292). Oxford: Pergamon Press.

Wright, B. D. (1999). Rasch Measurement Models. In G. N. Masters and J. P. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 85-97). Oxford: Pergamon.

Wright, D. (1999). Student Mobility: A Negligible and Confounded Influence on Student Achievement. *Journal of Educational Research*, *92* (6), 347-353.

Wright, N. K. and Dorans, N. J. (1993). *Using the Selection Variable for Matching or Equating*. Princeton, NJ: Educational Testing Service.

Yang, M., Goldstein, H., Rath, T. and Hill, N. (1999). The Use of Assessment Data for School Improvement Purposes. *Oxford Review of Education*, *25* (4), 469-483.

Yeung, A. S. and Marsh, H. W. (1997). *Gender Differences in the Development of English and Maths Constructs: Longitudinal Models of Academic Self-concept and Achievement*. Paper Presented at the Annual Conference of the Australian Association for Research in Education (AARE). 30 November - 4 December 1997. Brisbane.

Young, D. and Fraser, B. (1993). Socioeconomic and Gender Effects on Science Achievement: An Australian Perspective. *School Effectiveness and School Improvement, 4* (4), 265-89.

Young, D. J. (1998a). *Characteristics of Effective Rural Schools: A Longitudinal Study of Western Australian Rural High School Students*. Paper Presented at the Annual Meeting of the American Educational Research Association. April 13-17, 1998. San Diego, CA.

Young, D. J. (1998b). Rural and Urban Differences in Student Achievement in Science and Mathematics: A Multilevel Analysis. *School Effectiveness & School Improvement, 9* (4), 386-418.

14

Appendices

Appendix 14.1
Number of Grades 3 and 5 students who have zero and perfect scores in the 1995 to 2000 BST numeracy and literacy tests

i) Grade 3

		Numeracy		Literacy	
Occasion	N	Zero	Perfect	Zero	Perfect
1995	10,283	32	153	20	64
1996	11,095	46	100	33	23
1997	12,437	33	23	31	43
1998	12,794	36	96	31	41
1999	12,550	20	16	34	0
2000	12,677	39	42	41	6

ii) Grade 5

		Numeracy		Literacy	
Occasion	N	Zero	Perfect	Zero	Perfect
1995	10,735	20	23	6	7
1996	11,613	33	0	21	4
1997	11,973	24	11	22	7
1998	12,471	32	22	35	2
1999	12,976	37	19	12	3
2000	12,818	36	29	11	0

Notes:

N - Number of students participating

Appendix 14.2
Reliability estimates for the two-level and three-level unconditional models

Two-level models

	Reliability Estimate					
	Model-X		**Model-Y**		**Model-Z**	
Random Level-1 Coefficient	Numeracy	Literacy	Numeracy	Literacy	Numeracy	Literacy
INTRCPT1	0.452	0.429	0.374	0.362	0.371	0.366
SEX	×××	×××	0.143	0.125	0.143	0.091
AGE	0.190	0.176	0.164	0.113	0.161	0.088
ATSI	0.201	0.227	0.243	0.288	0.226	0.293
HOME	0.165	0.105	×××	0.177	×××	0.131
INOZ	×××	×××	0.060	0.068	0.062	0.08
TRANS			0.179	0.244		
YEARLEVL	0.612	0.550				
Y3NSCORE (or Y3LSCORE)			0.379	0.365	0.393	0.364

Three-level models

	Reliability Estimates					
	Model-X		**Model-Y**		**Model-Z**	
	Numeracy	Literacy	Numeracy	Literacy	Numeracy	Literacy
Random Level-1 coefficient						
INTRCPT1	0.482	0.453	0.629	0.621	0.372	0.416
ATSI	0.239	0.258	×××	×××	0.236	0.332
YEARLEVL	0.581	0.489				
TRANS			0.153	0.221		
Y3NSCORE or **Y3LSCORE**			0.317	0.316	0.339	0.330
Random level-2 coefficient						
INTRCPT1/INTRCPT2	0.745	0.888	0.820	0.986	0.824	0.987
YEARLEVL/INTRCPT2	0.924	0.974				

Notes: ××× - Variable has its random effect specified as fixed in this model
Shade - Variable not available for examination in the model

Appendix 14.3
Reliability estimates from null, Type A, Type B and Type C effects models

				Numeracy		Literacy	
	Model	**Effect**		Tran$^{\propto}$	Non-Tran$^{\beta}$	Tran$^{\propto}$	Non-Tran$^{\beta}$
Level-1	Null^{c}	Intercept	INTRCPT1, P0	0.333	0.323	0.332	0.314
	Type A	Intercept	INTRCPT1, P0	0.444	0.439	0.477	0.467
		Prior$^{\lambda}$	Y3SCORE, P1	0.289	0.278	0.270	0.252
	Type B	Intercept	INTRCPT1, P0	0.440	0.438	0.477	0.467
		Prior$^{\lambda}$	Y3SCORE, P1	0.291	0.276	0.280	0.257
	Type C	Intercept	INTRCPT1, P0	0.440	0.439	0.476	0.466
		Prior$^{\lambda}$	Y3SCORE, P1	0.291	0.277	0.279	0.257
Level-2	Null^{c}	Stable	INTRCPT1/ INTRCPT2,	0.859	0.836	0.844	0.825
		Change	INTRCPT1/ OCC,	0.092	0.053	0.106	0.116
	Type A	Stable	INTRCPT1/ INTRCPT2,	0.635	0.602	0.533	0.459
		Change	INTRCPT1/ OCC,	0.178	0.173	0.193	0.170
		Gender	SEX/ INTRCPT2,	0.136	0.111	0.093	0.068
		Age	AGE/ INTRCPT2,	×××	0.074	×××	×××
		Race	ATSI/ INTRCPT2,	×××	×××	0.281	0.270
		Transience	TRANS/ INTRCPT2,	0.143		×××	
		Prior$^{\lambda}$	Y3SCORE/ INTRCPT2,	×××	×××	0.124	0.083
	Type B	Stable	INTRCPT1/ INTRCPT2,	0.435	0.445	0.403	0.372
		Change	INTRCPT1/ OCC,	0.176	0.171	0.162	0.145
		Gender	SEX/ INTRCPT2,	0.132	0.122	0.082	0.057
		Transience	TRANS/ INTRCPT2,	0.107		×××	
		Prior$^{\lambda}$	Y3SCORE/ INTRCPT2,	×××	×××	0.066	0.065
	Type C	Stable	INTRCPT1/ INTRCPT2,	0.433	0.442	0.402	0.366
		Change	INTRCPT1/ OCC,	0.162	0.153	0.153	0.133
		Gender	SEX/ INTRCPT2,	0.131	0.121	0.079	0.057
		Transience	TRANS/ INTRCPT2,	0.111		×××	
		Prior$^{\lambda}$	Y3SCORE/ INTRCPT2,	×××	×××	0.067	0.069

Notes:
$\propto$ - Using all the students matched (Schools = 482)
β - Using only those students matched in the same school (Schools = 479)
c - Simplest longitudinal model
λ - Prior achievement; that is, Y3NSCORE or Y3LSCORE
××× - Error term deleted from the model
Shade - Variable (TRANS) not available for examination in this model

Appendix 14.4 Final estimation of variance components using null model, Type A, Type B, Type C effects models for numeracy

					Transience Data Set (School = 482)				**Non-Transience Data Set** (Schools = 479)			
	Model	**Effect**			Var. Comp.	df	Chi-Square	P-value	Var. Comp.	df	Chi-Square	P-value
Level-1	Null[c]	INRTCPT1,		R0	0.293	889	2092.408	0.000	0.229	865	2043.068	0.000
&		leve1-1,		E	0.998				0.986			
Level-2	Type A	INTRCPT1,		R0	0.036	855	2429.197	0.000	0.041	827	2390.683	0.000
		Y3NSCORE slope,		R6	0.011	1802	2607.009	0.000	0.012	1768	2476.946	0.000
		level-1,		E	0.558				0.539			
	Type B	INTRCPT1,		R0	0.036	853	2815.163	0.000	0.041	827	2664.416	0.000
		Y3NSCORE slope,		R6	0.011	1801	2602.994	0.000	0.012	1768	2477.277	0.000
		level-1,		E	0.557				0.539			
	Type C	INTRCPT1,		R0	0.036	853	2829.700	0.000	0.041	827	2681.985	0.000
		Y3NSCORE slope,		R6	0.011	1801	2602.653	0.000	0.012	1768	2477.048	0.000
		level-1,		E	0.557				0.539			
Level-3	Null[c]	INTRCPT1	/INTRCPT2,	U00	0.246	467	3992.676	0.000	0.218	463	3415.422	0.000
		INTRCPT1	/OCC,	U01	0.009	467	593.553	0.000	0.017	463	528.843	0.018
	Type A	INTRCPT1	/INTRCPT2,	U00	0.045	462	1369.678	0.000	0.041	446	1184.674	0.000
		INTRCPT1	/OCC,	U01	0.004	462	594.425	0.000	0.004	446	563.110	0.000
		SEX	/INTRCPT2,	U10	0.005	462	548.168	0.004	0.004	446	499.957	0.039
		AGE	/INTRCPT2,	U20	×××	×××	×××	×××	0.006	446	478.017	0.143
		TRANS	/INTRCPT2,	U30	0.012	462	543.460	0.005				
	Type B	INTRCPT1	/INTRCPT2,	U00	0.018	459	842.302	0.000	0.020	460	867.615	0.000
		INTRCPT1	/OCC,	U03	0.004	462	593.000	0.000	0.004	462	595.073	0.000
		SEX	/INTRCPT2,	U10	0.005	461	546.635	0.004	0.005	462	525.939	0.021
		TRANS	/INTRCPT2,	U20	0.009	460	523.247	0.022				
	Type C	INTRCPT1	/INTRCPT2,	U00	0.017	456	841.379	0.000	0.020	457	862.577	0.000
		INTRCPT1	/OCC,	U03	0.003	461	584.793	0.000	0.003	461	586.890	0.000
		SEX	/INTRCPT2,	U10	0.005	461	546.688	0.004	0.005	462	525.860	0.021
		TRANS	/INTRCPT2,	U20	0.009	460	523.538	0.021				

Notes: Shade - The variable TRANS is not available for examination in this model
××× - Error term deleted from the model
c - Simplest longitudinal model

Appendix 14.5 Final estimation of variance components using null model, Type A, Type B, Type C effects models for literacy

					Transience Data Set (School = 482)				**Non-Transience Data Set** (Schools			
	Model	**Effect**			Var. Comp.	df	Chi-Square	P-value	Var. Comp.	df	Chi-Square	P-value
Level-1	Null	INTRCPT,		R0	0.228	889	2078.405	0.000	0.208	865	2053.243	0.000
&		level-1,		E	0.999				0.975			
Level-2	Type A	INTRCPT1,		R0	0.034	856	2709.206	0.000	0.037	829	2715.071	0.000
		Y3LSCORE slope,		R7	0.008	1332	2565.146	0.000	0.008	1302	2441.730	0.000
		level-1,		E	0.449				0.430			
	Type B	INTRCPT1,		R0	0.034	856	2926.802	0.000	0.037	829	2870.664	0.000
		Y3LSCORE slope,		R7	0.008	1332	2604.834	0.000	0.008	1302	2446.625	0.000
		level-1,		E	0.450				0.431			
	Type C	INTRCPT1,		R0	0.034	856	2938.313	0.000	0.037	829	2890.214	0.000
		Y3LSCORE slope,		R7	0.008	1332	2603.480	0.000	0.008	1302	2440.800	0.000
		level-1,		E	0.450				0.431			
Level-3	Null	INTRCPT1	/INTRCPT2,	U00	0.216	467	3624.594	0.000	0.185	463	3188.469	0.000
		INTRCPT1	/OCC,	U01	0.011	467	580.362	0.000	0.013	463	572.901	0.001
	Type A	INTRCPT1	/INTRCPT2,	U00	0.027	358	755.303	0.000	0.022	335	593.746	0.000
		INTRCPT1	/OCC,	U01	0.003	358	440.141	0.002	0.003	335	391.363	0.018
		SEX	/INTRCPT2,	U10	0.002	358	427.898	0.007	0.002	335	344.721	0.345
		ATSI	/INTRCPT2,	U30	0.077	358	409.449	0.031	0.087	335	399.146	0.009
		Y3LSCORE	/INTRCPT2,	U60	0.001	358	423.365	0.010	0.001	335	363.837	0.134
	Type B	INTRCPT1	/INTRCPT2,	U00	0.014	462	816.965	0.000	0.013	460	740.112	0.000
		INTRCPT1	/OCC,	U01	0.003	466	582.199	0.000	0.003	462	570.557	0.001
		SEX	/INTRCPT2,	U10	0.002	466	528.071	0.024	0.002	462	477.661	0.297
		Y3LSCORE	/INTRCPT2,	U70	0.001	466	519.329	0.044	0.001	462	521.198	0.029
	Type C	INTRCPT1	/INTRCPT2,	U00	0.014	460	817.544	0.000	0.013	457	737.365	0.000
		INTRCPT1	/OCC,	U01	0.003	465	579.336	0.000	0.003	461	565.698	0.001
		SEX	/INTRCPT2,	U10	0.002	466	528.013	0.024	0.002	462	477.548	0.299
		Y3LSCORE	/INTRCPT2,	U70	0.001	466	519.486	0.043	0.001	462	521.282	0.029

Note: Null - Simplest longitudinal model

Appendix 14.6
Exploring the relationship between stable and change school effects

This appendix reports on the analyses carried out to examine why the correlations between the stable and change school effects for numeracy are positive while the corresponding correlations for literacy are negative (results in Table 9.11, Chapter 9). In particular, this appendix reports on the analyses undertaken to examine the following:

(a) the differences in score distribution between the two outcome measures;

(b) the direction of the correlation coefficient between the stable school effects and the change school effects in the null model;

(c) the direction of the correlation coefficient between the stable school effects and the change school effects when MLwiN (Browne et al., 2001) software is employed instead of the HLM5/3L (Raudenbush et al., 2000) software; and

(d) the direction of the correlation coefficient between the stable school effects and the change school effects if the literacy test is broken down into its sub-tests, that is, language and reading.

Comparison of distribution of scores

One plausible explanation of the contradictory results presented in Table 9.11 (Chapter 9) is the existence of a ceiling in the level of achievement in literacy at Grade 5 and the absence of such a ceiling for numeracy. It would seem unreasonable to assume that primary schools stop teaching their Grade 5 students once the students reach a competence level of achievement in literacy. However, at the secondary school level, Thorndike (1973a) argued that reading tests were testing reasoning, and that schools had stopped teaching reading. Thorndike based his argument on the proposition that performance in reading, after the basic decoding skills are mastered, is primarily an indicator of the general level of the individual's thinking and reasoning processes rather than a set of distinct and specialized skills. It is possible that the better students at Grade 5 level have mastered the basic decoding skills, and that little is being done in schools to develop reading skills and it is likely that Thorndike's argument could hold for the better students at Grade 5.

It is reasonable to assume that if indeed the ceiling exists, then it must stem from the tests used to measure achievement in literacy at Grade 5. Because, the tests are designed to identify the low achievers (Hungi, 1997), this means that the levels of achievement of the more capable students could be of less concern to the test developers especially in literacy. Consequently, it could be that most good schools have reached the ceiling as set by the test developers and, therefore, any further progress made by the schools can not be measured using the literacy tests. If this is the case, then using these tests, the progress made by the poor schools could still be measured because their levels of performance in literacy are below the ceiling set by the test developers. However, this does not mean that the poor schools can not genuinely catch up with the good schools without the existence of a ceiling effect.

It should be noted that the Rasch scaling approach employed in this study seeks to avoid (by using the logistic transformation) the ceiling effect that makes it difficult to observe much change when the raw scores of the student approach perfect scores. However, a ceiling effect could still be there because in Rasch scaling the abilities of

the students with near perfect raw scores are generally estimated with larger errors compared to the abilities of the students with average (or near average) raw scores. Moreover, a ceiling effect could have been introduced into the scores because an approximation procedure was employed to estimate the ability scores of the students with perfect scores. Hence, although the numbers of students who obtained perfect scores on each testing occasion were small (see Appendix 14.1), the possibility of a ceiling effect needs to be examined.

Obviously, for the above argument to hold it would mean that the frequency distributions of the two outcome measures must differ noticeably. In particular, the frequency distribution of the literacy score of the students at Grade 5 would have to be clearly negatively skewed compared to that of their numeracy scores. In addition, the frequency distributions of the literacy scores for the four cohorts of students plotted on the same graph should provide some evidence of increasing skewness in the distribution of the scores with the successive cohorts of students to reflect the ceiling effect.

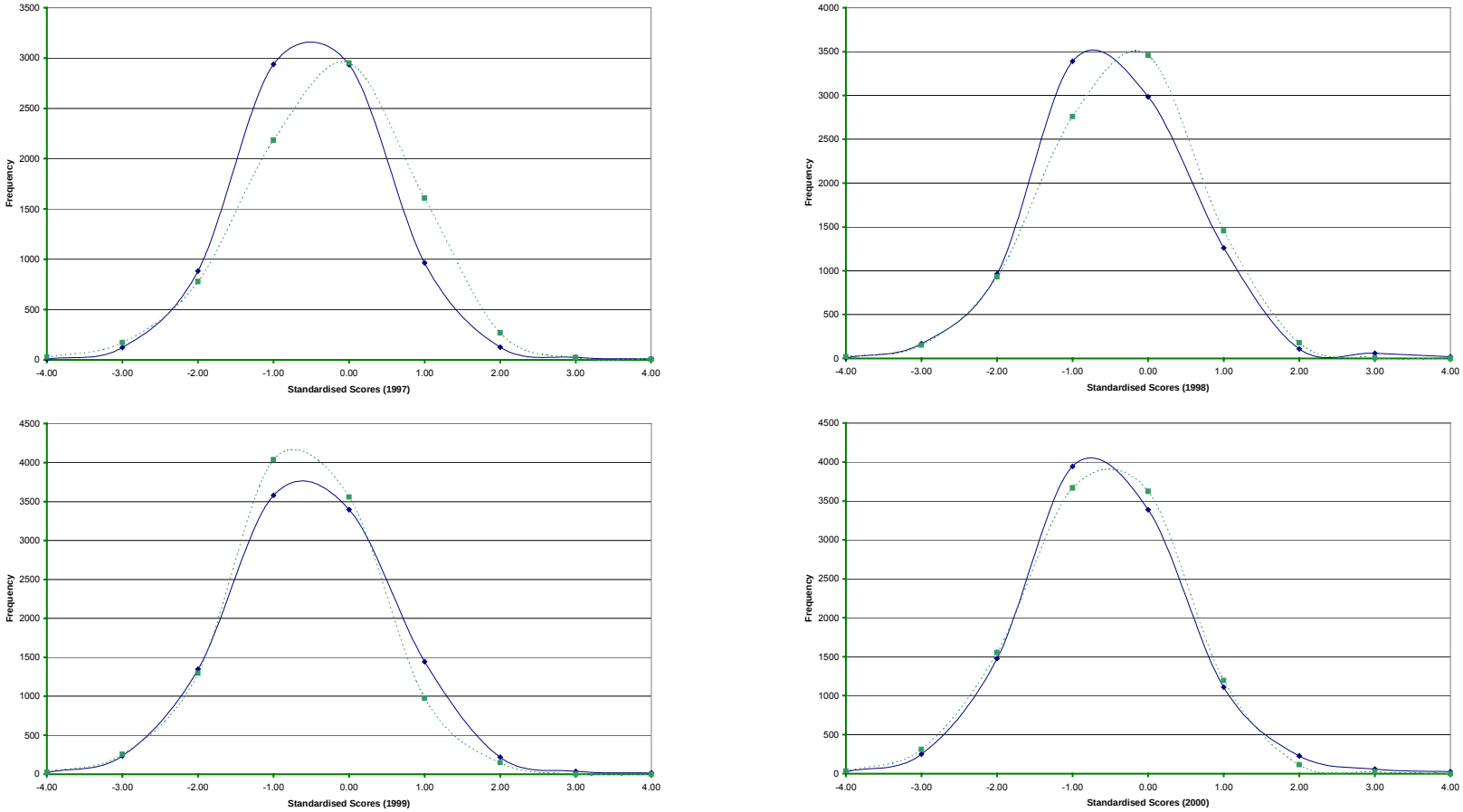
Figures A9.5 and A9.6 show frequency distribution plots of standardized numeracy and literacy scores for the four cohorts of students involved in this study at Grade 5. These plots were obtained using the data that includes all the students who could be matched (N=37,832). The standardization of the scores for each outcome measure was done separately. However, the standardization of the scores from all the four occasions was carried out simultaneously for each outcome measure.

Figure A14.1 displays the frequency distributions of the numeracy and literacy scores for each cohort of students on four separate plots, that is, one plot for numeracy and literacy for each cohort of students. Figure A14.2 displays the frequency distributions of the two outcome measures for the four cohorts of students on two separate plots, that is, one plot for numeracy and the other for literacy. In Figure A14.2, the percentages of students in each category of scores are used in the graphs rather than the raw frequencies to enable direct comparison of plots within each outcome measure. It is not easy to compare the plots using raw frequencies because of the differences in participation rates in the tests and, consequently, the different numbers of students who could be matched from each testing occasion.

For parsimony, all students with scores equal to or greater than four standardized scores have been placed in the same category in the plots shown in Figures A14.1 and A14.2. Likewise, all students with scores equal to or less than four standardized scores have been placed in one category.

In Figure A14.1 it is clear that the distributions of scores for the two outcome measures are consistently similar which seems to suggest that the chances of a ceiling effect for one of the outcome variables is unlikely. And from Figure A14.2 it is evident that within the same outcome measure, with only small variations, the distributions of the scores for the four cohorts of students follow a similar pattern and they are all near the normal distribution.

However, from Figure A14.2 it appears that although the plots for literacy are nearly similar they are nevertheless more separated compared to those of numeracy. This indicates that, compared to the distribution of the numeracy scores, the distribution of the literacy scores differs slightly for the four cohorts of students. These differences are investigated next.



Note: In all the plots the solid line is for numeracy and the dotted line is for literacy

Figure A14.1 A comparison between the frequency distribution of the numeracy scores and the literacy scores at Grade 5

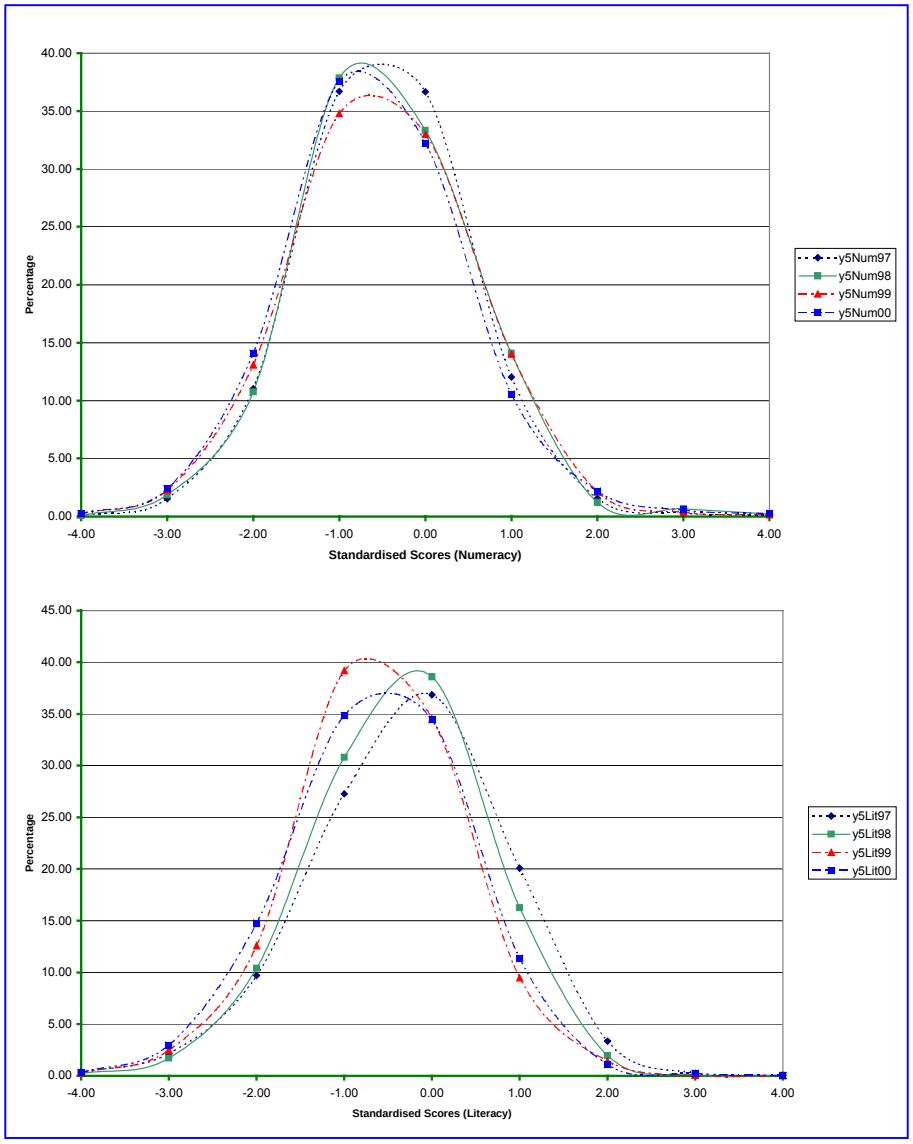


Figure A14.2 Frequency distributions of the numeracy and literacy for the four cohorts of student

Table A14.1 presents the values of skewness of the distributions of the numeracy scores as well as literacy scores for the four cohorts of students while Figure A14.3 shows graphical plots of the skewness values. The two broken lines in Figure A14.3 show the linear trends in the skewness of the plots with successive cohorts of students for each outcome measure.

In interpreting the plots in Figure A14.3 it should be considered that the scale used on the skewness axis exaggerates the differences in skewness of the plots over time. Despite the exaggeration, the plots in Figure A14.3 show that little changes have occurred in the skewness of the distributions of the numeracy scores compared to the changes that have occurred in the skewness of the distribution of literacy scores. In

particular, the figure shows that skewness of the distributions for literacy scores has increased over time while it has decreased marginally (or remained almost constant) for numeracy scores. However, the actual values (in Table A14.1) indicate that the changes in skewness values of the literacy plots are small. It should be noted that the slope of the linear trend line for skewness of the distributions of the literacy scores is 0.09, and for the numeracy scores it is –0.03. Therefore, although Figure A14.3 indicates that the skewness of distributions of the literacy scores has increased over time, the evidence may, nonetheless, not be sufficient for making sound conclusions regarding the existence of a ceiling effect in the literacy tests.

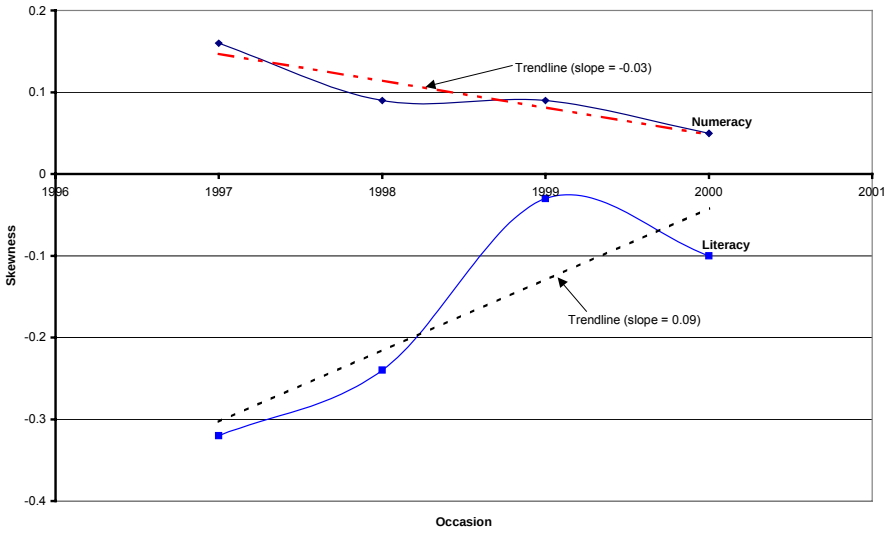


Figure A14.3 Trends in skewness of distributions of numeracy and literacy scores

Table A14.1 Skewness of the distribution of numeracy and literacy scores at Grade 5

	1997	1998	1999	2000
Numeracy	0.16	0.09	0.09	0.05
Literacy	-0.32	-0.24	-0.03	-0.10

Correlation between stable and change school effects in the simplest longitudinal models

Another plausible explanation to the results presented in Table 9.11 (Chapter 9) could lie in the differences in the contribution made by the student background characteristics to each of the two outcome measures. It is likely that the two outcome measures provide similar relationships between the stable school effects and the change school effects when the student background characteristics are not included in the model and the relationships only differ when the contribution made by these characteristics are taken into account. If this is found to be the case, then it would be reasonable to conclude that the contradictory results presented in Table 9.11 arise

entirely from differences in the nature of contribution made by the student background factors to each of the two outcome measures.

In order to test whether or not the control for the student background factors is the cause of the discrepancy in the numeracy and literacy results, the simplest longitudinal model for each of the two outcome measures (Y5NSCORE and Y5LSCORE) is estimated using the two data sets. In order to test if the observed discrepancy also exists at Grade 3, the simplest longitudinal model is also estimated with the Grade 3 scores as the outcome measures, that is, Y3NSCORE for numeracy and Y3LSCORE for literacy.

The analyses in this section are undertaken using the two leading computer programs in multilevel modelling, that is, HLM5/3L (Raudenbush et al., 2000) and MLwiN (Browne et al., 2001). The aim here is to check if similar results are obtained using the two programs. In particular, employing both HLM5/3L and MLwiN would ascertain whether or not the discrepancy in the numeracy and literacy results is due to the analytical approach employed by HLM5/3L.

For HLM5/3L, the simplest longitudinal model is the same as Equation 9.18 presented in Chapter 9, that is, the model has the time trend variable OCC as the only predictor and no other predictor variables are specified at any level of this model. The equation for this model is presented again below.

Level-1 model

$$\mathbf{Y}_{itj} = \boldsymbol{\pi}_{0tj} + \mathbf{e}_{itj}$$

Level-2 model

$$\boldsymbol{\pi}_{0tj} = \boldsymbol{\beta}_{00j} + \boldsymbol{\beta}_{01j}\mathbf{OCC}_{tj} + \mathbf{r}_{0tj}$$

Level-3 model

$$\boldsymbol{\beta}_{00j} = \boldsymbol{\gamma}_{000} + \mathbf{u}_{00j}$$

$$\boldsymbol{\beta}_{01j} = \boldsymbol{\gamma}_{010} + \mathbf{u}_{01j}$$

Equation A9.1

All the components in Equation A9.1 carry the same meaning as described in Chapter 9 for models for Type A and Type B effects. However, the outcome measure, Y, can either be the achievement (Rasch score) of the student at Grade 5 or at Grade 3 depending on the model being estimated. As in the analyses presented in Chapter 9, the time trend variable OCC in Equation A9.1 is group-mean centred in these analyses (Kreft, 1995; Kreft et al., 1995).

For MLwiN, the simplest longitudinal model is as follows:

$$\mathbf{Y}_{itj} = \boldsymbol{\beta}_{0itj} + \boldsymbol{\beta}_{1j}\mathbf{OCC}_{tj}$$

where:

$$\boldsymbol{\beta}_{0itj} = \boldsymbol{\beta}_{0} + \mathbf{v}_{0j} + \mathbf{u}_{0tj} + \mathbf{e}_{0itj}$$

$$\boldsymbol{\beta}_{1j} = \boldsymbol{\beta}_{1} + \mathbf{v}_{1j}$$

Equation A9.2

and where:

$\mathbf{Y}_{itj}$ is the achievement (Rasch score) of student *i* in school *j* at occasion *t*;

$\boldsymbol{\beta}_{0}$ is the grand mean;

β_{1j} is the OCC slope for school j;

u_{0ij} is the residual for occasion;

v_{0j} is residual for school;

v_{1j} is residual for OCC slope at Level-3; and

e_{0itj} is residual for student.

As for the case of HLM5/3L analyses, the time trend variable OCC in MLwiN analyses is group-mean centred.

Table A14.2 shows the correlations between the stable school effects and the change school effects obtained using the two data sets for each of the outcome measures and using HLM5/3L and MLwiN. The first panel in Table A14.2 present the correlations obtained when Grade 5 scores are used as the outcome measures while the second panel presents the correlations when Grade 3 scores are used as the outcome measures.

For the Grade 5 scores, the results in Table A14.2 confirm that the relationships between the stable school effects and the change school effects differ for the two outcome measures even without the control for student background characteristics. As found above in the models for estimation of Type A and Type B effects, the relationship is positive for numeracy and negative for literacy regardless of the data set used and regardless of the computer software employed in the analyses.

Table A14.2 Correlations between stable and change school effects from the simplest longitudinal model

		Transience$^{\propto}$		**Non-Transience**$^{\beta}$	
Outcome		HLM	MLwiN	HLM	MLwiN
Grade 5 Score**					
	Numeracy	0.92	0.93	0.99	0.99
	Literacy	-0.44	-0.47	-0.53	-0.58
Grade 3 Score**					
	Numeracy	0.36	0.34	0.31	0.29
	Literacy	0.94	0.94	0.71	0.68

Notes: $\propto$ - Using all the students matched (Schools = 482).
β - Using only those students matched in the same school (Schools = 479)
** - All the correlations are significant at the 0.01 level.

With the Grade 3 scores as the outcome variables, the results in Table A14.2 indicate that the relationships between the stable school effects and the change school effects for literacy are in the same direction as for numeracy. Again, the results here are consistent regardless of the data set used and the computer program employed to analyze the data. For numeracy, the correlations at Grade 5 are overwhelmingly extremely strong ($\geq$0.92), which is contrary to the correlations at Grade 3 (0.29 to 0.36), indicating that some changes have occurred to the relationships between the stable school effects and the change school effects across the two grades. For literacy, interestingly, the correlations are positive at Grade 3 (0.68 to 0.94) while they are negative at Grade 5 (–0.44 to -0.58) indicating a major shift in the relationships between the stable school effects and the change school effects across the two grades.

Figures A9.8 and A9.9 show Level-3 (school-level) plots of residuals and their ranks generated by MLwiN (Browne et al., 2001) software following the estimation of the simplest longitudinal models for numeracy and literacy with the Grade 5 scores as the

outcome variable and using the non-transience data set. The first panel of each figure displays the plot of the stable school effects (intercept) while second panel shows the change school effects. In order to illustrated the relationship between the two components, the residuals for three schools (labelled 'A', 'B', and 'C') are highlighted in both figures. The three schools are selected on the basis of their stable school effect numeracy residuals: school A has the lowest residual (thus, least effective); school B has residual near zero; and school C has the highest residual (thus, most effective).

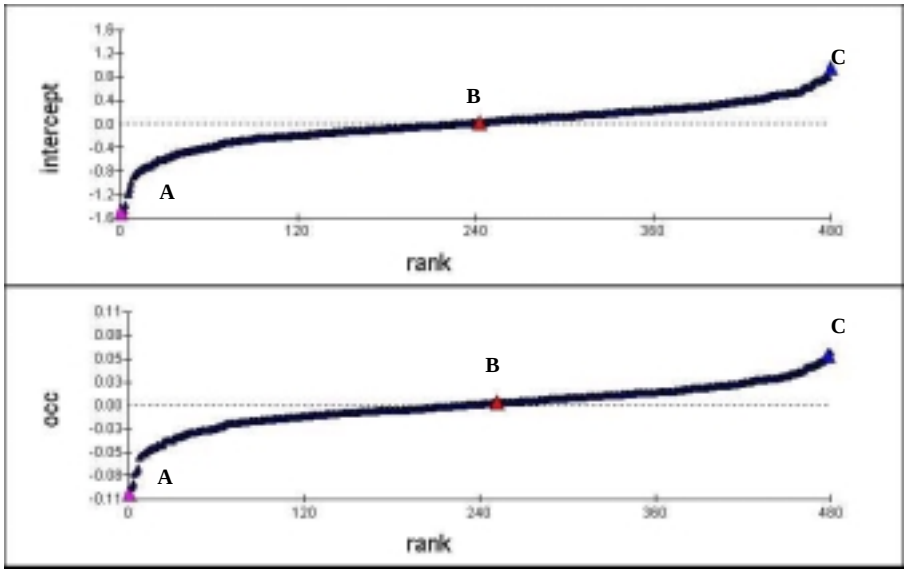


Figure A14.4 Level-3 residual plots for numeracy

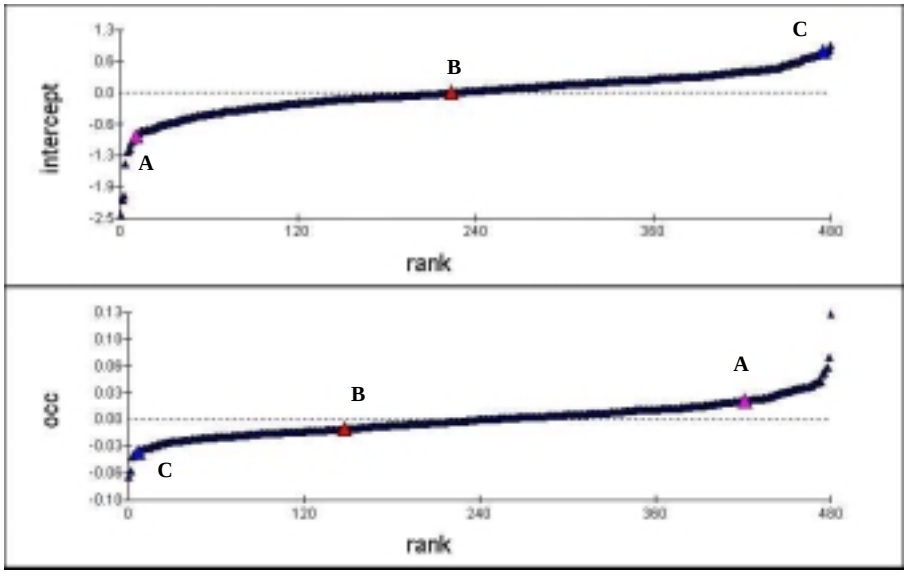


Figure A14.5 Level-3 residual plots for literacy

From Figure A14.4, it can be observed that the three schools retain their relative position in the OCC plot, that is, school A has the lowest change effect residual (thus, gained the least), school B has near a zero residual and school C the has the highest residuals (thus, gained the most). Thus, Figure A14.4 illustrates that a school that shows more than expected average performance in numeracy also shows more than expected increase in performance in numeracy over time, and vice versa.

However, from Figure A14.5 it can be observed that schools A and C have changed their relative positions in the OCC plot. Thus, Figure A14.5 illustrates that a school that shows more than expected average performance in literacy shows less than expected increase in performance in literacy over time. The figure also illustrates that a school that shows less than expected average performance in literacy shows more than expected increase in performance in literacy over time.

The first panels of Figure A14.4 and A14.5 show that the three schools maintain their relative positions across the two outcome measures. Thus, schools that show more than expected average performance in numeracy also show more than expected average performance in literacy. Alternatively, schools that show less than expected average performance in numeracy also show less than expected average performance in literacy. However, schools A and B are not the schools with the extreme intercept residual as can be observed from Figure A14.5. Thus, the association between the stable school effects across the two outcome measures is positive, but it is by no means perfect.

Correlations between stable and change school effects for reading and language

Another interesting facet to investigate is the relationship between the stable school effects and the change school effects with literacy scale broken into its sub-scales, that is, reading and language. Hungi (1997) examined the factor structure of the Basic Skills Tests and found strong evidence to support the existence of (a) a numeracy factor and not clearly separate number, measurement, and space factors, and (b) a literacy factor and clearly separate language and reading factors. Hence, it is possible that the problem lies in the lack of unidimensionality of the literacy tests. This can be examined by analysing the two sub-scales of literacy separately to compare the correlations between the stable school effects and the change school effects in the two sub-scales. The correlations from the two sub-scales can also be compared with the correlation obtained when the two sub-scales are pooled to form one scale, that is, the literacy scale. If the directions of correlations between the components of school effects in either reading or language (or both) differ from the direction obtained in the literacy scale, then it will be reasonable to conclude that the contradictory results presented in Table 9.11 (in Chapter 9) arise from a lack of unidimensionality of the literacy test.

In order to test whether or not the dimensionality of the literacy test is the problem, all the Grade 3 and Grade 5 literacy tests from the six occasions (1995 to 2000) are equated to construct common scales: one for Language and the other for Reading. The same procedure and techniques employed to equate the numeracy and literacy tests (described in Chapter 6) are employed here to equate the language and the reading tests. After successful equating of the tests, language and reading scores are computed for all the individuals who could be matched (N= 37,832) for Grade 3 and for Grade 5. Finally, HLM5/3L is employed to estimate the simplest longitudinal model for language and reading using the Grade 5 scores as the outcome variables and then using the Grade 3 scores as the outcome variables.

Table A14.3 presents the correlations between the stable school effects and the change school effects using the two data sets for language and reading. For ease of comparisons, the correlation obtained when the two sub-scales are combined to form a single literacy scale (result in Table A14.2) are also presented in this table. The first panel in Table A14.3 displays the correlations obtained when Grade 5 scores are used as the outcome measures while the second panel displays the correlations when Grade 3 scores are used as the outcome measures.

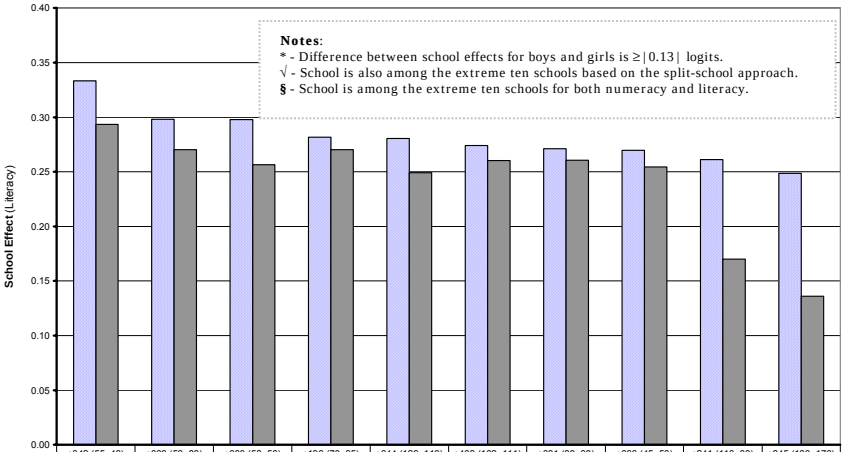
Thus, the results in Table A14.3 indicate that the relationships between the stable school effects and the change school effects for both language and reading do not differ in direction from what was obtained when the two scales are combined to form a single literacy scale. It should be noted that at the Grade 3 level, all the correlation coefficients are positive while at the Grade 5 level they are negative regardless of the variable used as the outcome. This indicates that the shift in the nature of the relationships between the stable school effects and the change school effects across the two grades exists with or without combining the language and reading sub-scales to make a single literacy scale. Thus, it appears that the problem does not lie in the lack of unidimensionality of the literacy test.

Table A14.3 Correlations between stable and change component of school effect for language, reading and literacy

Outcome		Tran$^{\propto}$	Non-Tran$^{\beta}$
Grade 5 Score**			
	Language	-0.60	-0.54
	Reading	-0.25	-0.36
	Literacy	-0.44	-0.53
Grade 3 Score**			
	Language	0.99	0.98
	Reading	0.99	0.99
	Literacy	0.94	0.71

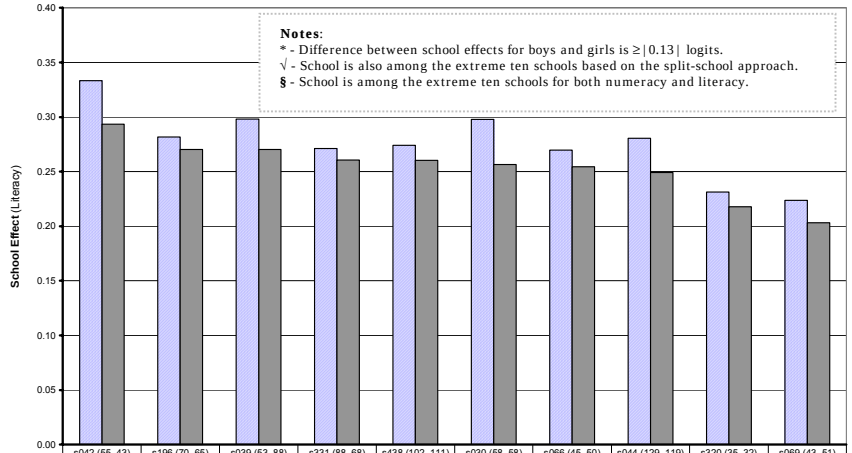
Notes: $\propto$ - Using all the students matched (Schools = 482).
β - Using only those students matched in the same school (Schools = 479)
** - All the correlations are significant at the 0.01 level.

Appendix 14.7
Histogram plots for the most and least effective schools in literacy by gender



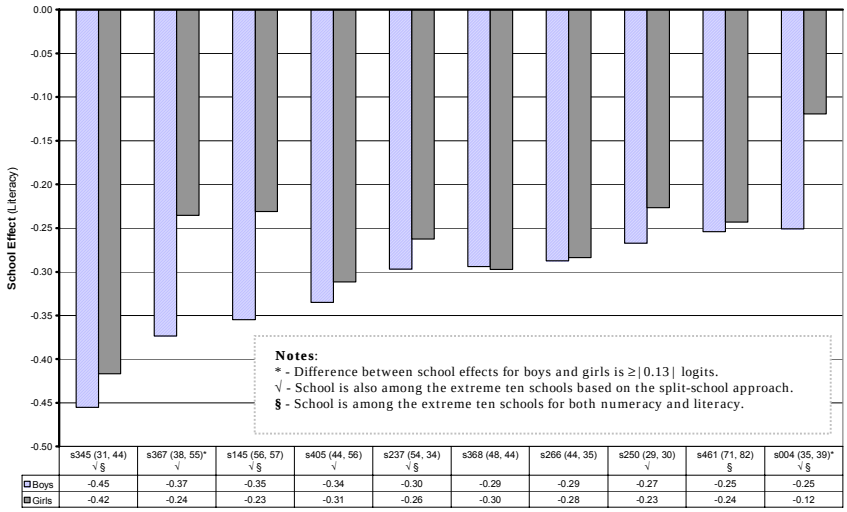
	s042 (55, 43) √	s039 (53, 88) √ §	s030 (58, 58) √ §	s196 (70, 65) §	s044 (129, 119) √ §	s438 (102, 111) √	s331 (88, 68) √ §	s066 (45, 50)	s241 (110, 88) √	s245 (183, 176) √ §
Boys	0.33	0.30	0.30	0.28	0.28	0.27	0.27	0.27	0.26	0.25
Girls	0.29	0.27	0.26	0.27	0.25	0.26	0.26	0.25	0.17	0.14

Figure A14.6 Top ten effective schools for boys in literacy



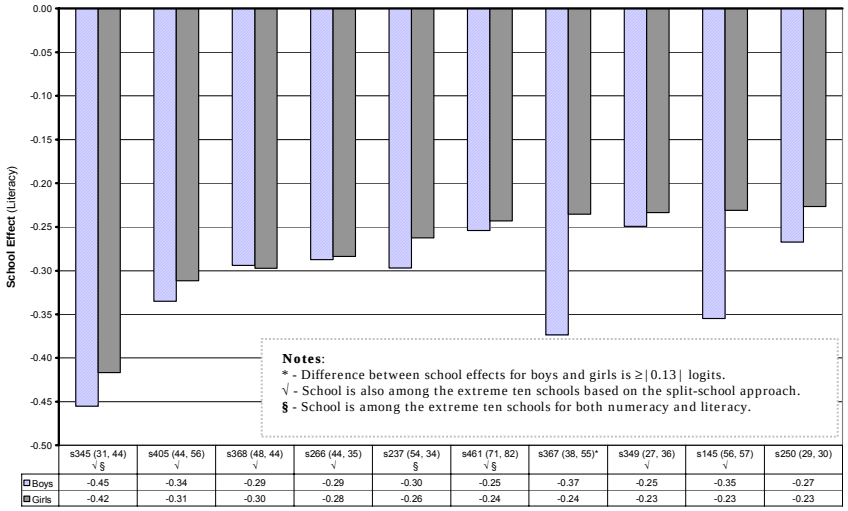
	s042 (55, 43) √ §	s196 (70, 65) √ §	s039 (53, 88) √	s331 (88, 68) √ §	s438 (102, 111) √	s030 (58, 58) §	s066 (45, 50)	s044 (129, 119) √ §	s320 (35, 32) √	s069 (43, 51) √
Boys	0.33	0.28	0.30	0.27	0.27	0.30	0.27	0.28	0.23	0.22
Girls	0.29	0.27	0.27	0.26	0.26	0.26	0.25	0.25	0.22	0.20

Figure A14.7 Top ten effective schools for girls in literacy



	s345 (31, 44) √ §	s367 (38, 55)* √	s145 (56, 57) √ §	s405 (44, 56) √	s237 (54, 34) √ §	s368 (48, 44)	s266 (44, 35)	s250 (29, 30) √	s461 (71, 82) §	s004 (35, 39)* √ §
Boys	-0.45	-0.37	-0.35	-0.34	-0.30	-0.29	-0.29	-0.27	-0.25	-0.25
Girls	-0.42	-0.24	-0.23	-0.31	-0.26	-0.30	-0.28	-0.23	-0.24	-0.12

Figure A14.8 Ten least effective schools for boys in literacy



	s345 (31, 44) √ §	s405 (44, 56) √	s368 (48, 44) √	s266 (44, 35) √	s237 (54, 34) §	s461 (71, 82) √ §	s367 (38, 55)*	s349 (27, 36) √	s145 (56, 57) √	s250 (29, 30)
Boys	-0.45	-0.34	-0.29	-0.29	-0.30	-0.25	-0.37	-0.25	-0.35	-0.27
Girls	-0.42	-0.31	-0.30	-0.28	-0.26	-0.24	-0.24	-0.23	-0.23	-0.23

Figure A14.9 Ten least effective schools for girls in literacy

Shannon Research Press
Adelaide, South Australia
ISBN: 1-920736-02-6