THE VICTORIAN QUALITY SCHOOLS PROJECT:
A STUDY OF SCHOOL AND TEACHER EFFECTIVENESS

REPORT (Volume 1)
Report to the Australian Research Council
Re: Grant File No. A79531244

Chief Investigators:
Professor Peter W. Hill
Kenneth J. Rowe
Philip Holmes-Smith

Associate Investigator:
Dr V. Jean Russell

1996

Centre for Applied Educational Research
Faculty of Education, The University of Melbourne
Preface

This report follows a successful application for funding from the Australian Research Council (ARC) to complete additional stages of the Victorian Quality Schools Project (VQSP) – a longitudinal study of school and teacher effectiveness undertaken in a large sample of government, Catholic and independent schools in Victoria. Commencing in 1992, the project had been jointly funded by the Victorian Directorate of School Education (DSE) and The University of Melbourne, under an agreement that expired at the end of 1994. Funding existed to complete the third and final phase of data collection during 1994 and to complete analyses of the data for the purposes of reporting to participating schools and to the DSE. However, funding was unavailable to pursue important findings arising from preliminary analyses of the 1992 and 1993 data through further analysis of the 1994 data and the combined 1992-1994 data, or to complete intensive on-site case studies of selected schools designed to illuminate results derived from statistical modelling of the data.

By the end of 1994, the VQSP had generated one of the most comprehensive and significant data bases on school and teacher effectiveness ever assembled. It is the first major empirical study of school effectiveness undertaken within Australia, and one of few internationally to collect longitudinal measures of schools, teachers and students using a design that has enabled the estimation of both school and class/teacher effects. Preliminary analyses of the data (Hill et al., 1993a; Rowe et al., 1994, 1995; Hill & Rowe, 1996) have suggested that conventional views on school-level effects need to be re-evaluated in the light of the VQSP findings, and that much of what is generally regarded as effects due to student and school characteristics are in fact due to effects operating at the class/teacher-level. Further analysis and qualitative follow-up of this data base offered the very real prospect of major conceptual and methodological advances in the field of school effectiveness research, with consequent practical spin-offs for school reform endeavours world-wide.

Thus, support from the ARC was sought to fund an additional one-year phase of the project during the course of 1995 and early 1996 to examine in greater detail both quantitative and qualitative hypotheses arising from initial analyses of the VQSP data. Funding granted by the ARC via Grant No. 79531244 has enabled this additional work to be undertaken. The present report describes this work and presents the relevant findings. A key feature of the report is the documentation of the data-gathering instruments that have been used throughout the project and their related coding manuals, together with the normative and measurement properties of the scales and constituent items. A summative overview of the major findings from the study is also provided.
Acknowledgments

The willing participation of the students, parents, teachers and school principals who contributed to this project is gratefully acknowledged. The work of Judith Clements and Margaret Cowin in their capacity as field investigators for the multi-site case studies is much appreciated, as is the support provided by Tim Jones, Office Manager of the Centre for Applied Educational Research. The successful completion of this project owes much to their joint efforts.

The support of senior officers of the Directorate of School Education (Victoria) is also appreciated, together with that provided by the Catholic Education Office (Victoria) and the Association of Independent Schools of Victoria. The project was initiated by the State Board of Education in 1992 and continued with full funding support until December 1994 from the Victorian Directorate of School Education under an agreement with The University of Melbourne. The additional funding provided by the Australian Research Council during 1995 under Grant No. 79531244, is acknowledged with thanks.

Thanks is also due to the former Vice Chancellor of The University of Melbourne, Professor David Penington, for his role in establishing the Centre for Applied Educational Research in January 1993; to Professor Kwong Lee Dow, Dean of the Faculty of Education, for his on-going guidance and counsel; to Professors Brian Caldwell and Hedley Beare, and the administrative staff of the Faculty of Education for their encouragement and valued assistance. We likewise appreciate the collaborative support from colleagues in the Departments of Psychology and Paediatrics at the University of Melbourne, together with that from colleagues at the Australian Council for Educational Research. We also value collaborative exchange with colleagues at the University of London Institute of Education and with those in the Faculty of Educational Science and Technology at the University of Twente, the Netherlands.

Thanks are due to Peter Hart, Department of Psychology, University of Melbourne for kind permission to use the “Work Environment” items from the School Organizational Health Questionnaire.
# Contents (Volume I)

Preface ........................................... i  
Acknowledgments ................................. ii  
Context for the Study ............................ 1  
Problems endemic to studies of *educational effectiveness* 2  
1. Lack of empirical evidence .................. 2  
2. Lack of longitudinal studies modelling change 3  
3. The issue of data-analytic methods  .......... 3  
4. Problems in the choice of outcome measures 5  
5. The issue of sample size ..................... 9  
6. Statistical adjustments ..................... 9  
Focus of the report ............................. 11  

Aims of the Study ............................... 12  

Study Design and Method ........................ 13  
Sample and design ............................. 13  
Measures ........................................ 14  
Analytic methods .............................. 22  

Summary of Results ............................. 31  

Qualitative Case Study ........................ 68  
Purpose ......................................... 68  
Research questions ............................ 68  
Sample and methods ............................ 69  
Procedure ...................................... 71  
Data analyses ................................... 71  
Results ......................................... 72  
Conclusion ...................................... 74  
Appendix 1. Research question 5: Analyses and results 74  

Concluding Comments ........................... 83  

References ....................................... 87
Context for the Study

The provision of schooling is one of the most massive and ubiquitous undertakings of the modern state. Schools account for a substantial proportion of public and private expenditure and are universally regarded as vital instruments of social and economic policy aimed at promoting individual fulfilment, social progress and national prosperity. It therefore comes as no surprise that there has long been intense interest in knowing how effective the provision of schooling is and how it can be improved. As noted by Hill, Holmes-Smith and Rowe (1993a, p. 1): “This is an especially sensitive issue at the present time given the level of consensus regarding the importance of school education as an element of micro-economic reform and in meeting the constantly changing demands of the modern workplace”. What is surprising, however, is the shakiness of our knowledge about educational effectiveness in terms of outcomes for students, teachers and schools. Even more intriguing is that the journey taken by researchers since the 1960’s in search of answers appears, a quarter of a century later, to be approaching a point close to where the journey began. For recent accounts of this journey see Bosker and Scheerens (1994); Creemers (1994a); Creemers and Scheerens (1994); Mortimore, (1991, 1992); Reynolds and Cuttance (1992); Reynolds, Teddlie, Creemers et al. (1994); Sammons (1996); Scheerens (1992, 1993).

Despite the difficulties entailed in defining an effective school (see Cheng, 1996; Mortimore, 1991; Sammons, 1996), the work on school effectiveness to date has primarily focussed on the search for ways to measure the quality of a school. Although the term quality is likewise problematic (see Istance & Lowe, 1991), the “...measurement of the quality of schooling is of critical importance at a time when so much school reform in so many parts of the world is being undertaken” (Mortimore, 1991, p. 214). Concern about the quality of school education has become a high priority policy issue in all OECD countries (see OECD, 1989) where attention has focused on ways of assessing the quality of schools (typically defined and articulated in terms of academic attainment), of identifying factors associated with effective schooling, and on using such knowledge to achieve further improvements in quality (Banks, 1992; Chapman, Angus, Burke & Williamson 1991; Coleman & Collinge, 1991; Creemers & Scheerens, 1989; Cuttance, 1992a; McGaw, Piper, Banks & Evans, 1992; Reynolds & Cuttance, 1992).

It is noted frequently (eg., Creemers & Scheerens, 1994) that school effectiveness research grew out of studies of educational effectiveness focussing on production functions (Fraser, Walberg, Welch & Hattie, 1987; Hanushek, 1979; Monk, 1992), and more especially out of the initial sociologically oriented input-output studies by Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld, and York (1966), and subsequently by Jencks, Smith, Acland, Bane, Cohen, Gentsis, Heynes and Michelson (1972). These researchers were interested primarily in issues of equity and the influence of the school relative to that of sociologically-determined background characteristics of students. Their findings were interpreted as casting serious doubts on the capacity of schools to make a difference relative to the influence of the socio-economic capital of the home. Reynolds, Hargreaves and Blackstone (1980) summarised this interpretation in the following terms: “...variations in what children learn at school depends largely upon

---

1 Mortimore (1991, p. 216) suggests the following simple definition: “An effective school is one in which pupils progress further than might be expected from consideration of its intake".

Centre for Applied Educational Research, The University of Melbourne
variations in what they bring and not on variations in what schools offer them” (p. 208). In what has become a familiar pattern, the conclusions arrived at by this early research were consistent with the prevailing climate of social and political opinion.

Early studies of school effectiveness such as those by Brookover, Beady, Flood, Schweitzer and Wisenbaker (1979); Edmonds (1979a) and by Rutter, Maughan, Mortimer, Ouston and Smith (1979), were conceived largely as a reaction to the Coleman and Jencks conclusions. The Brookover, Edmonds and Rutter studies adopted a different starting point and focussed on identifying contextual features of schools in which students were performing better than their counterparts in comparable schools, after adjusting for the effects of intake characteristics. Given this starting point, the positive conclusions from such studies and the enthusiasm with which they were promoted was not unexpected.

The message from this work has been that effective schools are characterised by an ethos or culture oriented towards learning expressed in terms of high standards and expectations of students, an emphasis on basic skills, a high level of involvement in decision-making and professionalism among teachers, cohesiveness, clear policies on matters such as homework and student behaviour, and so on. Moreover, effective schools are also characterised by outstanding educational leadership, particularly as exercised by the principal and directed at establishing agreed goals, increasing competence and involvement of staff and at clarifying roles and expectations. Edmonds (1979b) was the first to summarise these features into what has become known as the “five factor model” of school effectiveness, namely: purposeful educational leadership, challenging teaching and high expectations of students, involvement of and consistency among teachers, a positive and orderly climate, and frequent evaluation of student progress.

This “five factor model” forms the basis of what might be termed the optimistic account of school effectiveness research – an account that presents a positive picture of the role and efficacy of contextual school influences. In addition to the well known critiques of the “five-factor model” (eg., Creemers, 1994a; Ralph & Fenessey, 1983; Scheerens & Creemers, 1989), there are several problems with the optimistic account, not the least of which is that it has built upon an extremely fragile research base. For example, Rutter et al.’s (1979) study was based on observations made in just twelve inner London schools. Nonetheless, there are at least six major problems with this optimistic account that are highlighted here. It should be noted, however, that these ‘problems’ are also endemic to the general field of research concerned with educational effectiveness.

Problems endemic to studies of educational effectiveness

1. Lack of empirical evidence

The most widely cited summarisations of school characteristics as alterable correlates of educational effectiveness have been provided by Edmonds (1979a, 1979b, 1981), Tomlinson (1980), and Purkey & Smith (1983). More recent summaries have been provided by Austin & Reynolds (1990) and Levine & Lezotte (1990). Common to each of these summaries, is the tendency to produce recipes for “effectiveness”, based on disparate and often anecdotal ‘findings’ reported in the literature, rather than on
empirical research evidence. Sadly, the little empirical evidence available to date about school effectiveness *per se* and *educational effectiveness* in general, is not extensive with most of the knowledge base being derived from small-scale case studies; but mostly from scholarly reviews and comment (e.g., Good & Weinstein, 1986; Levine & Lezotte, 1990; Wilson & Corcoran, 1988). In fact, Creemers (1994a) notes: "Until recently, the state of the art in school effectiveness research was aptly described by the fact that the research reviews outnumbered the total of empirical investigations" (p. 9).

The dangers of relying predominantly on anecdotal recipes and scholarly comment is that their constituent rhetoric have a strong influence on major educational policy initiatives. However, the research base and related empirical evidence to support such policy initiatives is extremely limited. On the basis of an intensive study of models of school effectiveness, Banks (1992, p. 199) observes:

> Research on effective schools is being used to shape major policy-making initiatives in Australia and overseas, even though what makes some schools more effective than others remains an open question. Because clear and unequivocal messages to educators and policy makers are yet to emerge from the research, unquestioning acceptance of the current findings should be a cause for concern.

2. Lack of longitudinal studies modelling change

This lack of empirical evidence is underscored by the fact that there have been relatively few large-scale studies capable of providing sustainable generalisations about educational effectiveness, and fewer still that have collected longitudinal data that are essential for examining factors affecting *growth* or *change*. To date, most quantitative studies of school effectiveness have been restricted to cross-sectional designs, or have collected data at two time points only. Modelling *effectiveness* or *change* in such circumstances typically involves fitting conditional models in which measures of students' learning outcomes (adjusted for, or conditional on, intake characteristics and prior achievement) are regressed on explanatory variables. Because such studies are invariably non experimental, the drawing of inferences about *change* is problematic. In addition, measures of change based on only one or two time points are notoriously unreliable (Bryk & Raudenbush, 1987, Goldstein, 1979, 1986a; Kessler & Greenberg, 1981; Raudenbush, 1989; Raudenbush & Bryk, 1988; Rogosa & Willett, 1985; Willett, 1988). Nuttall, Goldstein, Prosser and Rasbash (1989) suggest that it is necessary to be cautious in interpreting "...any study of school effectiveness that relies on measures of outcome in just a single year, or of just a single cohort of students. Long time series are essential for a proper study of stability over time" (p. 775). While the advice is apt, the logistical problems in mounting and maintaining such studies entail severe practical constrains, resulting in a virtual absence of studies conducted over long periods of time.

3. The issue of data-analytic methods

The methods typically used to analyse the derived data from quantitative studies of educational effectiveness have not allowed for the modelling of complex inter-relationships between inputs, processes and outcomes, including indirect effects and reciprocal effects; nor have they taken into account the inherent nested structure of schooling and the typical organization of students into classes taught by particular
teachers. These are problems that only relatively recent methodological advances have addressed.

Two developments are especially worthy of comment. The first is the development of structural equation modelling techniques that enable the testing of complex inter-relationships among variables within a framework that takes into account measurement error (Bentler, 1980, 1989; Jöreskog & Sörbom, 1979, 1993a; McDonald, 1978; Muthén, 1984). The second is the development of multilevel analysis techniques that account for the hierarchical sampling structure of the data and enable simultaneous estimation of the influence of variables operating at different levels of analysis (Aitkin & Longford, 1986; Bock, 1989, Bryk & Raudenbush, 1988, 1992; Goldstein, 1986, 1987, 1995; McDonald, 1994; Rasbash & Woodhouse, 1996; Raudenbush & Willms, 1991; Rowe & Hill, in press).

It is well documented (eg., Bosker, Creemers & Scheerens, 1994; Goldstein, 1995; Hill & Rowe, 1996; Raudenbush & Willms, 1991, 1995; Rowe, Hill & Holmes-Smith, 1995) that studies of educational effectiveness in terms of estimating the effects of schooling on student learning over time "...share two key features: the fact that student growth is the object of inquiry, and the fact that such growth occurs in organizational settings" (Raudenbush & Bryk, 1988, p. 424). Raudenbush and Bryk go on to note that these features correspond, in turn, to two of the most troublesome and enduring methodological problems in educational research, namely: the problem of measuring change (Bryk & Raudenbush, 1987; Goldstein, 1979, 1986a; Harris, 1963; Rogosa & Willett, 1985; Willett, 1988) and the problem of analysing multilevel data (Bryk & Raudenbush, 1992; Goldstein, 1986b, 1987, 1995; Raudenbush, 1989).

Unfortunately (as noted by Bosker & Scheerens, 1989, 1994; Hill & Rowe, 1996; Raudenbush & Willms, 1991; Rowe & Hill, 1994; Rowe et al., 1995; Scheerens, 1992), relatively few studies have been undertaken that have accounted for the inherent nested or multilevel organizational structure of schooling with students grouped into classes and taught by particular teachers, despite mounting evidence for the importance of instructional effects at the class/teacher-level (Creemers, 1994b; Schaffer, Nesselrodt, & Stringfield, 1994; Teddlie, 1994). For example, Monk (1992) cites a number of studies in support of the observation that: "One of the recurring and most compelling findings within the corpus of production function research is the demonstration that how much a student learns depends on the identity of the classroom to which that student is assigned" (p. 320). In similar vein, Raudenbush and Willms (1991, p. xi) observe:

An irony in the history of quantitative studies of schooling has been the failure of researchers' analytic models to reflect adequately the social organisation of life in classrooms and schools. The experiences that children share within school settings and the effects of these experiences on their development might be seen as the basic material of educational research; yet until recently, few studies have explicitly taken account of the effects of particular classrooms and schools in which students and teachers share membership.

Although studies designed to provide estimates of student-, class- and school-level effects are few in number, they have nevertheless prompted a major reassessment of knowledge about educational effectiveness. For example, in a review of the relevant literature, Reynolds and Packer (1992, p. 173) conclude:

Centre for Applied Educational Research, The University of Melbourne
On the causes of school effects, it seems that early beliefs that school influences were distinct from teacher or classroom influences were misplaced, since a large number of studies utilizing multi-level modelling show that the great majority of variation between schools is in fact due to classroom variation and that the unique variance due to the influence of the school, and not the classroom, shrinks to very small levels.

To avoid, or at least minimise these problems, Rowe et al. (1995) recommend that: (1) the design of educational performance and school effectiveness studies should be longitudinal, with repeated measures on multiple cohorts of students nested within classes and schools to estimate change over time, and (2) that multilevel analysis be employed to account for the inherent hierarchical structure of the data (i.e., repeated measures clustered within students who are grouped within classes and schools). Similarly, Gray et al. (1995, pp. 99-100) and Goldstein (1996, p. 6) stipulate four criteria for what is suggested are minimum requirements for satisfactory inference in studies of educational effectiveness (see also Stringfield, 1994a). These requirements are:

• that studies be longitudinal, covering at least three data-collection periods, and contain measures of outcomes and prior achievement on individual students;
• that the data analysed be available on a minimum of three cohorts of students, and preferably more;
• that proper multilevel statistical analyses be undertaken to ensure the validity of statistical inferences; and
• that there be a demonstrated commitment to an examination of the data for systematic changes over time.

It should be noted, however, that this is not a criticism of all existing studies of school effectiveness. Together with Goldstein (1996), we would argue that many of these studies have made important contributions to the clarification of these requirements. Neither do we suggest that educational policy makers, school administrators or teachers should forgo the adoption of both policies and practices based on the best available evidence. Rather, our purpose in emphasising these requirements in quantitative studies of educational effectiveness is to highlight present and future possibilities in the statistical modelling of data on change or progress in student learning outcomes.

4. Problems in the choice of outcome measures

Whereas the long-term goals of school education may be expressed as the enhancement of young peoples' access to and participation in society, as well as preparation for meeting the constantly changing demands of the modern workplace (OECD, 1983, 1986), the most direct and readily accessible measures of schooling outcomes are obtained from assessments of students' academic attainments and progress. Herein, however, lies a dilemma that is evidenced in strident critiques of traditional and prevailing psychometric models for test and examination modes of assessment (e.g. Berlak, 1992) and an equally strident chorus of concern for the deleterious effects of test-driven and test-dominated curricula (e.g. Kellaghan, Madaus & Airasian, 1982), or as Frederiksen (1984) expressed it aptly: "the tail of testing that wags the curriculum dog" (p. 201). Nisbet (1993, p. 25) highlights this dilemma as follows:
In today's schools, assessment is a main influence on how pupils learn and how teachers teach. Whether assessment is in the form of examinations and tests, or marks and grades for coursework, its influence is pervasive. Often it distorts the process of learning through teaching to the test, cramming, short-term memorising, anxiety and stress - to the extent that learning to cope with assessment has become almost as important as the genuine learning which such assessments are supposed to measure. For many young people, assessment dominates education.

To date, for the purposes of determining educational and occupational access, standards monitoring, accountability, and school effectiveness, the measurement of learning outcomes at the student, school, system, national and international levels has relied almost exclusively on the use of standardised achievement tests or public examinations (Goldstein & Lewis, 1996; Haney, Madaus & Lyons, 1993; Scheerens, 1993, 1995; Sutherland, 1996). Robert Wood's (1986) comment that, "In Britain, to talk of educational measurement is to talk of examinations" (p. 197), continues to hold as it has for the past fifty years. In the United States, especially, the use of standardised achievement tests, dominated by the psychometric technology of item response modelling, has prevailed over the same period (Goldstein, 1996a; Goldstein & Wood, 1989; Madaus & Raczek, 1996; Wood, 1987a,b). Although the use of such tests and examinations for the measurement and evaluation of educational effectiveness is typically justified on the grounds of maximising reliability and ensuring comparability, it is argued cogently that this has been mostly at the expense of validity (Broadfoot, 1994a; Lacey & Lawton, 1981; Moss, 1994). For example, in summarising the British, European and North American attempts at curriculum and assessment reform during the 1970's, Lacey and Lawton (1981, pp. 229-230) warned:

...conventional standardised achievement tests have inherent risks as instruments of evaluation for accountability since they seldom cover more than the common core or very basic curriculum units. Thus, as the sole instrument, they may be highly deceptive because of lacking content validity. ...test scores as such have low information value about the outlying processes as well as the environmental and administrative frame conditions necessary to understand and appreciate the skills and efforts needed to fulfil a certain educational goal.

Problems of content validity, however, would appear to be less acute in studies that have made use of public examination results, such as the study reported by Tymms (1993), since public examinations are designed to assess learning outcomes as set out in some detail in syllabi, from which it can be assumed that teachers and schools have followed closely. Where examinations have been used as outcome measures, differences between classes and faculties within schools are typically large and substantially greater than differences among schools, although effects are not especially consistent across faculties or from year to year.

In several school effectiveness studies, the problem of the content validity of items contained in standardised tests has been recognised and attempts have been made to assess the extent to which students have had the opportunity to learn the content represented in individual test items. Where this has been done, it has been observed that 'opportunity to learn' is a major explanation for patterns of performance on the tests. This is true in the case of the study reported by Bosker et al. (1990) and has been a consistent finding in various international studies of achievement (see Bosker & Scheerens, 1989, 1994). Unfortunately, it is not always possible to determine whether
lack of 'opportunity to learn' reflects unsatisfactory test validity or inadequate instruction.

From the U.S. there has long been criticism of the utility of standardised tests as measures of either learning or competence (Darling-Hammond, 1994; Frederiksen, 1984; Kellaghan et al., 1982; Linn, 1981, 1986; Newmann & Archbald, 1990; Wigdor & Garner, 1982). Newmann and Archbald (1990) argue, for example, that "...most data currently used to assess schools' performance, especially scores on standardised tests, fail to measure meaningful forms of human competence and that significantly new forms of assessment need to be developed" (p.164). From Britain, the use of traditional tests and examinations has likewise been the focus of intense critical discussion and calls for alternatives (see Broadfoot, 1984, 1986a,b, 1988, 1994a,b, 1996; Gipps, 1990, 1994a,b,c; Gipps & Murphy, 1994; Horton, 1990; Murphy & Torrence, 1991; Murphy & Broadfoot, 1995; Wolf, 1995). Broadfoot (1994a, p. 5) observes: "...dissatisfaction with traditional testing approaches is now widespread, explicit and clearly articulated".

For a variety of epistemological and methodological reasons such criticism has been manifestly ignored by researchers in almost all large-scale studies of factors affecting student learning outcomes. This is most notable in school effectiveness research (e.g. Bosker, Creemers & Scheerens, 1994; Gray & Wilcox, 1995; Mortimore, 1995; Mortimore, Sammons & Thomas, 1994; OECD, 1989; Raudenbush & Willms, 1991; Reynolds et al., 1994), and in related methodological commentaries (e.g., Draper, 1995; Goldstein, 1996c, Morris, 1995; Raudenbush & Willms, 1995), where the identification of 'effective schools' continues to be made on the basis of limited operational definitions of what it means to be a 'good school'. Applying an apparent embargo on measures of presumably desirable affective and behavioural outcomes of schooling, the most common approach in these studies is to identify those schools with aggregated scores on standardised tests of reading and mathematics (or on examinations) that are higher than average, after making adjustments for measures of student intake characteristics - regardless of the curriculum validity of those tests. In so doing, there seems to be little awareness that "...the majority of such tests assess skills in terms of generalised academic abilities and enduring cognitive 'traits' rather than specific learning outcomes arising from classroom instruction" (Hill & Rowe, 1996, p. 7). Under such circumstances, claims for 'school effectiveness' per se, are at best, tenuous.

Elements of this criticism have gained credence in the areas of standards monitoring and performance assessment in the U.S. where alternative approaches to obtaining more curriculum-specific and "authentic" (Wiggins, 1989) measures are being considered (e.g. Floden, 1994; Lesh & Lamon, 1992; Moss, 1994; Murphy, 1990; Nitko, 1995; O'Connor, 1992; Resnick & Resnick, 1992, Shavelson, 1994; Taylor, 1994). Similarly, the work of Broadfoot (1994a,b), Butterfield (1995), Drummond (1993), Harlen (1994), Gipps (1994a,b,c), Gipps & Murphy (1994), Torrence and Pryor (1995), and Wolf (1995) is representative of a growing number of U.K. educationalists who have called into question prevailing 'mechanistic', 'objectives-driven' modes of assessment and re-assert the teacher's professional role in education by challenging the widespread assumption that teachers' assessments are less reliable than those obtained from examinations and tests.

In response to such concerns, many large education systems, particularly those in Australia, Britain, Canada, Europe, New Zealand and the United States that have had
student monitoring procedures in place for many years, have within the last ten years, embarked on the process of dismantling and re-building their assessment programmes, to make them more inclusive of the broad range of competencies required of students, and to improve their utility in terms of reporting to students, parents, employers, and so on. One approach to monitoring that runs parallel to these reforms in assessment is the development of profiles as frameworks for recording, reporting and monitoring students' educational progress\(^2\) – an approach in which Australia has an established record.\(^3\)

Nonetheless, relative to the extensive consideration typically given to the selection of explanatory variables, quantitative studies of school effectiveness have typically paid scant attention to the choice of outcome measures. Yet choice of outcome measures has major implications for the conclusions that one might draw regarding the impact of student-, class- and school-level effects (see Hill & Rowe, 1996). With the exception of several UK studies, such as the above-mentioned study by Tymms (1993) that have made extensive use of results for secondary students in subject-based public examinations, almost all studies of school effectiveness have used norm-referenced, standardised achievement tests as outcome measures.

To the extent that standardized tests are generally not sensitive to the specifics of what has actually been taught, school and class effect sizes are likely to be smaller (Bosker & Scheerens, 1989; Madaus, Kelleghan, Rakow, & King, 1979). Moreover, student achievement on such tests is likely to be more positively correlated with prior measures of general ability and home background factors than with measures of instructional characteristics of classrooms.

In short, as stressed by Hill and Rowe (1996), the choice of outcome measures in studies of school effectiveness requires careful attention since the use of different measures is likely to lead to different conclusions about educational effectiveness. In particular, it is likely that standardised tests, public examinations and performance tasks assess different kinds of learning outcomes and are not equally sensitive to student-, class- and school-level factors. This leads to the conclusion that for a full understanding of educational effectiveness researchers need to examine a range of student learning outcome measures and not assume that educational effectiveness can be adequately described or determined on the basis of any one kind of measure.

---

\(^2\) It is beyond the scope of this report to review the extensive literature related to profiles and records of achievement generated from within the United Kingdom during the 1970's and 1980's. For examples of this literature see: Balogh (1982), Baumgart (1986), Broadfoot (1982, 1984, 1986a, 1987), Mansell (1986), Rowntree (1977), SCRE (1977), Stevenson (1983).

\(^3\) Examples of this expertise are illustrated by Griffin (1990; Griffin & Nix, 1991; Griffin, Smith & Burrill, 1995a,b); by the Victorian Ministry of Education's Profile Handbooks (Victoria, 1990, 1991, 1992); by the Australian Council for Educational Research in the Victorian Achievement Studies (Adams, Doig & Rosier, 1991; McGaw et al., 1989), validation of the Mathematics Profiles (Adams, Barley & Wu, 1992) and in the Developmental Assessment Resource for Teachers (DART - Forster, 1994; Forster, Mendelovits & Masters, 1995; Masters & Forster, 1995a,b); the New South Wales Basic Skills Testing Program (Masters et al., 1990); the Western Australian Monitoring Standards in Education Program (Titmanus et al., 1993) and Student Outcome Statements (Western Australia, 1993); the Queensland Assessment of Student Performance (1990); and the recent Victorian Curriculum and Standards Framework (Board of Studies, 1995).
5. The issue of sample size

It is evident that research aimed at elucidating school- and class-level effects implies sampling designs that are sufficiently large to allow for the simultaneous estimation of effects at the school, class/teacher and student levels with stability (see Burstein, Kim & Delandshere, 1989; Gray, 1989; Goldstein, 1987, 1995). At the secondary level, departmental effects are also likely to be significant and should be built into the sampling design (Tymms, 1993). Studies that attempt to estimate variation at the school level with samples of less than thirty schools are likely to produce estimates that are unstable. In practice, oversampling and attention to sample maintenance at lower levels may also be necessary. For example, sample attrition is common in longitudinal studies, particularly within schools with high transience rates (Stringfield, 1994b). This can reduce significantly the number of cases at both the school and class levels and lead to biased estimates of variance at both these levels. In such cases, it may be advisable to undertake separate analyses that exclude schools and classes for which there have been high levels of student attrition, or seek to model the effects of transiency.

6. Statistical adjustments

Schools often vary significantly in the nature of their student intakes on account of socio-demographic factors and the existence of selective schools or private, fee-paying schools. In order to estimate the contribution of the school to student learning, studies of school effectiveness make use of statistical controls in an attempt to partial out the effects of such variability. As Bosker and Witzier (1995) observe, different studies employ very different kinds of statistical adjustments, although most can be categorised into one of the following three types:

*Unpredicted Achievement.* The most commonly used adjustment involves controlling for background characteristics of students such as general cognitive ability (as measured, for example, by verbal and/or non-verbal intelligence test scores) and family socio-economic status (e.g., using measures such as father’s occupation and family income as measured by entitlement to welfare payments or free lunches). The assumption underlying this kind of adjustment is that academic achievement at school is strongly influenced by sociological characteristics and that effectiveness is best conceived of as that part of achievement not explained by these characteristics.

*Learning Gain.* Another kind of adjustment involves controlling for achievement as measured on one or more previous occasions. By controlling for prior achievement, one is able to obtain estimates of learning gain over the intervening period or periods. For example, a number of British studies have been undertaken in which scores obtained in ‘A’-level examinations have been adjusted for scores obtained in GCSE examinations taken two years earlier. The assumption underlying this adjustment is that academic achievement at school is best predicted by prior achievement, which in turn reflects differences in student intake characteristics and any differential school effects prior to the initial achievement measure. Effectiveness is thus defined in terms of that part of achievement not predicted by prior achievement.

*Net Progress.* This involves a combination of both of the above kinds of adjustments, namely, controlling for both student background characteristics and prior achievement. This was the adjustment used in the above-mentioned study by Mortimore et al. (1988). The assumption underlying this adjustment is that achievement is significantly affected by
both prior achievement and intake characteristics and that effectiveness is best conceived as that part of achievement not explained by both of these factors.

While Bosker, Kremers and Lugthart (1990) provide evidence of a fairly high correlation between each of the three kinds of adjusted, or 'value-added' measures, it would appear likely that different adjustments will result in different estimates of the proportion of variance accounted for by school-, class- and student-level effects. This, in turn, suggests the importance of identifying clearly the particular kind of adjustment used, and where possible, to report results based on more than one kind of adjustment.

Regardless of which particular adjustment is employed, estimates of school-, class- and student-level effects will be influenced by under-specification (i.e., not measuring all relevant aspects of family background) and unreliability in measures of intake factors (resulting in attenuation of regression coefficients). The impact of having measures on only a small number of student background characteristics is difficult to predict because of the inter-correlated nature of such measures, although in many cases it is to be expected that it will lead to an under adjustment of achievement scores and to an over estimate of between-school and between-class differences. The impact of unreliability in intake measures invariably leads to over-estimates of the proportion of variance at the student level and to under-estimates of the effects at higher levels.

These methodological criticisms of the early school effectiveness research have provided the impetus for a relatively small number of 'second generation' studies and to an even smaller number of what Scheerens (1992, 1995) describes as 'state-of-the-art' studies (i.e., Bosker et al., 1990; Brandsma, 1993; Mortimore, Sammons, Stoll, Lewis & Ecob, 1988; Teddlie & Stringfield, 1993). These more recent studies consistently find that differences between schools, when relevant characteristics of students are fully taken into account, are important but not especially large — a finding that is confirmed by results from a comprehensive meta-analytic study by Bosker and Witziers (1995). Moreover, they are of an order of magnitude close to that estimated by Coleman and Jencks (i.e., ~ 9 per cent of the variance). At the same time, those studies that have been designed to enable the estimation of teacher and class-level effects have consistently identified larger proportions of between-class/teacher variance (see, for example, the ILEA Junior School Project reported by Mortimore et al., 1985, 1988, 1989; and especially the re-analysis of IEA data reported by Scheerens, Vermeulen & Pelgrum, 1989). This, in turn, has prompted a renewed focus on teacher or instructional effectiveness and to some re-definition of the fundamental questions underpinning educational effectiveness research (see Creemers, 1992, 1994; Slavin, 1994; Teddlie, 1994) and reflected in findings of recent large-scale, quantitative studies of educational effectiveness (see Reynolds, Sammons, Stoll & Barber, 1995; Scheerens & Creemers, 1995; Stringfield & Herman, 1995; Thomas, Sammons, Mortimore & Smees, 1995).

The small number of 'state of the art' educational effectiveness studies undoubtedly reflects the fact that the technical and logistical demands of such studies are immense. In the Australian context, the Victorian Quality Schools Project (described in this report) is the first major empirical study of school effectiveness, although there has been an important national study by McGaw and colleagues into parent and teacher perceptions of what makes an effective school (McGaw, Piper, Banks & Evans, 1992). For a recent
overview of work related to school effectiveness and school improvement in Australia, see Townsend (1995).

Focus of the report

This report summarises the aims, methods and key findings from three waves of data arising from the Victorian Quality Schools Project (VQSP) – a longitudinal, quantitative study of school and teacher effectiveness undertaken between 1992 and 1994 among government, Catholic and independent primary and secondary schools in Victoria. The report also provides an account of the work undertaken during the course of 1995 and early 1996, following a successful application for an Australian Research Council (ARC) Large Grant award. Specifically, support from the ARC was sought in 1994 to fund an additional phase of the project to pursue important findings arising from preliminary analyses of the 1992 and 1993 data through further analysis of the combined 1992-1994 data, and to complete intensive on-site case studies of selected schools designed to illuminate results derived from statistical modelling of the data. This additional work provided the research team with the opportunity to examine in greater detail both quantitative and qualitative hypotheses arising from initial analyses of the data.

Rather than present a detailed account of VQSP findings that are available elsewhere (see Hill et al., 1993a,b; Hill & Rowe, 1996, in press; Rowe & Hill, 1996, in press; Rowe et al., 1993, 1994, 1995), the present report provides a summative overview. A key element of the report, however, is the documentation of the data-gathering instruments that have been used throughout the project and their related coding manuals, together with the normative and measurement properties of the scales and constituent items. These are included in Appendices “A” to “C”. The raw data from the project are available upon request by interested researchers.

A major feature of the VQSP was the provision of extensive feedback to schools of their results, together with information allowing normative comparisons with other project schools. Participating schools responded with enthusiasm to this feedback (see Rowe et al., 1993). Indeed, many schools expressed an interest in receiving ongoing feedback beyond the life of the project. This prompted the idea of providing participating schools with the capacity to collect and analyse their own data with the assistance of specially written computer software.

Thus, in 1995, the data-gathering instruments designed and used for the third phase of the VQSP (see Appendix “C”) and for the second stage of the Schools of the Future - Co-operative Research Project (1995) were included in a software package known as the School Improvement Information Service (SIIS), developed by Hill, Rowe, Hill and Jones (1995). Written in Microsoft ACCESS™, the SIIS software was provided as a stand-alone custom application and designed to operate on IBM® compatible microcomputers with a 80386 processor or higher, requiring MS-DOS® version 6.0, or later, and Microsoft Windows® version 3.1 or later. The SIIS Users’ Guide is included in Appendix “D” of this report.
Aims of the Study

The Victorian Quality Schools Project (VQSP) was a research and development project designed to develop strategies for schools and school systems that would lead to improved outcomes for students, teachers and school leaders. The VQSP was conceived in the light of the international school effectiveness research literature and the limitations of that research, as noted above. In particular, the project built on the findings from an earlier, four-year longitudinal study of factors affecting students' reading achievement undertaken within the Victorian Ministry of Education, known as the 100 Schools Project - Literacy Program Study (Rowe, 1991a, 1991b, 1995). The VQSP was initiated as a project of the State Board of Education (Victoria), based on a formal proposal by Rowe (1991c), with the aim of seeking answers to the following key research questions:

1. What are the characteristics of schools in which students make rapid and sustained progress in Literacy (English) and in Mathematics, after adjusting for their intake factors and initial levels of achievement?

2. What are the characteristics of schools in which there are positive student attitudes and behaviours, positive perceptions by teachers of their work environment, and high levels of parent participation and satisfaction with their child's schooling?

In addition, the study aimed to contribute developmentally towards other aims, namely: to facilitate school improvement processes within participating schools, and to provide input to system-level quality assurance and accountability arrangements, policy development and planning.

The design of the study allowed for the exploration of inter-relationships among factors at three levels (student, teacher and school) over three time periods (1992, 1993 and 1994). The study design incorporates the following key features:

- a large sample of schools involving all students in each of five different Grade levels, thus ensuring generalisability of findings and enabling effects to be estimated at the level of the student, class/teacher and school;
- the use of multiple outcome measures, including students' achievements in English and mathematics, student attitudes and behaviours, teachers' affect levels and perceptions of their work environment, and parent participation in and satisfaction with their child's schooling;
- the use of Subject Profiles as a means of obtaining 'authentic', comprehensive and curriculum-specific teacher assessments of student achievement;
- the use of a longitudinal design to monitor students' progress and growth, as well as other changes within the school over time; and
- the use of analytic methods that allow exploration of interrelationships among factors at each level of analysis, as well as the simultaneous estimation of the

Initially, the study was entitled "The 150 Schools Project", but was subsequently changed to the "Victorian Quality Schools Project" from 1993 and onwards.

Centre for Applied Educational Research, The University of Melbourne
effects of factors at the student, class/teacher and school levels on rates of student progress, adjusting for student intake characteristics and initial achievement levels.

- qualitative case studies in six schools during 1995 to further explore issues identified through the quantitative analyses of the longitudinal data collected between 1992 and 1994.

**Study Design and Method**

**Sample and design**

Specific details of the sampling design of the study are given in Hill et al. (1993a), Hill and Rowe (1996), and Rowe et al. (1994). In brief, the study used a stratified probability sample of government and non-government elementary schools and secondary colleges. The sampling procedure adopted was a two-stage cluster design with schools being selected randomly but with probability proportional to their enrolment size at the first stage and entire cohorts of students in the selected schools at the second stage. Specifically, the design involved obtaining repeated measures on five, entire year-level cohorts of students from each participating school, initially in Grades: Preparatory (K), 2, 4, 7 and 9 (for 1992), and followed into subsequent Grades (for 1993 and 1994) as illustrated diagramatically below:

```
Year  Grade Levels
1992  K  2  4  7  9
1993  1  3  5  8  10
1994  2  4  6  9  11
```

Thus, over the three year duration of the study, student data were collected for each of the compulsory years of schooling.

Following written invitations, a total of 96 schools initially agreed to participate in the project. Useable data were received from 90 school sites in 1992, yielding a response rate of 88.5 per cent. For 1992, the achieved sample comprised 13,909 students, 931 teachers at 90 school sites; and for 1993: 12,630 students, 1110 teachers at 81 school sites – representing a student sample loss of approximately 10 per cent between 1992 and 1993. For 1994, data were obtained from 11,552 students and 1297 teachers from 73 school sites, representing a further 8.5 per cent reduction in the student sample. The reason for the increased numbers of teachers participating over the duration of the study was due to the encouragement by principals for wide participation by their teaching staff in completing the *Teacher Questionnaire* (see below), not just by those teachers directly concerned with students in the target year levels as shown above.

Sample attrition rates were higher than anticipated at the outset of the project, due to a number of factors, one of which was randomly missing data arising from a failure on the part of respondents to answer all questions. Moreover, similar to other longitudinal studies (see Stringfield, 1994), natural attrition played a part. Australia is characterised by a highly mobile society with high turnovers in school enrolments from year to year.
being common. In addition, state government policy changes throughout the three-year duration of the study resulted in the closure of approximately ten per cent of schools in the government sector which also had an impact on sample attrition. Finally, several schools withdrew from the project after the first year for a variety of reasons, but due mainly to alleged workload pressures on teachers.

**Measures**

The full data base for the *Victorian Quality Schools Project* contains a large number of variables relating to student background characteristics, achievement, behaviours and attitudes, classroom instruction; parent participation in and perceptions of various aspects of the schools attended by their children; teacher affect and perceptions of their work environment. Moreover, detailed data were collected related to aspects of perceptions and work undertaken by school leaders. The measures relevant to the three data collection phases are described briefly below. The variables measured in 1992 and 1993 are described in greater detail by Hill et al. (1993a, 1995) and by Rowe et al. (1993, 1994, 1995), and the data-gathering instruments and coding manuals are documented in Appendices “A” and “B” respectively. The 1994 data-gathering instruments, together with their related coding manuals, are recorded in Appendix “C”.

A problem which tends to afflict longitudinal studies of the kind described here, arises when the researchers wish to vary the data-gathering instrumentation during the course of a study. In the present study, as evidence accumulated during the first two years for very large variations in student achievements at the class/teacher-level (see Hill & Rowe, 1996), it was realised that in the initial design of the study insufficient attention had been given to instrumentation designed to measure key aspects of the classroom environment and especially instruction. This led to the development of a set of nine scales referred to below that were designed to elicit students’ opinions on various aspects of the teaching they experience. These scales were administered in 1994 only.

**Literacy and mathematics achievement**

In the design of the VQSP a conscious decision was taken to not depend primarily on standardized test results in measuring student achievement. This decision was not taken lightly, since most previous studies have made exclusive use of such measures, and (as noted earlier) rarely in the reports of these studies is reliance on standardized testing seen as problematic or requiring comment. The decision was made in response to serious reservations regarding the validity of available standardized tests as measures of the curriculum as taught in Victorian schools.

As already mentioned, issues related to the content validity of standardized tests have been the focus of much attention. In the context of undertaking large-scale studies of school effectiveness, Stringfield (1994) draws attention to the fact that so many standardized tests are "...either extremely broad in focus...and thus lacking in curriculum-related specificity, or so specific as to lack generalizability across schools or districts" (p. 57). Other writers have referred to the limited operational definition of effectiveness implicit in studies of school effectiveness that rely upon standardized tests to measure achievement. The consequences of such limited operational definitions may well be quite serious in terms of seeking to understand school effectiveness.
Thus, rather than embark on a project to develop batteries of standard assessment tasks covering all of the years of compulsory schooling, an alternative course of action was taken in the VQSP to obtain systematic teacher assessments of student progress in literacy (English) and mathematics using specially developed rating scales or Subject Profiles specifically designed to reflect the curriculum of Victorian schools, namely, the English Profiles (Victoria, 1991) and the Mathematics Profiles (Victoria, 1992). These Subject Profiles are ordered sets of indicators describing observable learning behaviours that have been empirically calibrated onto common measurement scales, or developmental growth continua, using a partial credit item response model (Adams & Khoo, 1992; Masters, 1982).

A full account of the development of these Subject Profiles is given by Griffin (1990), Griffin and Nix (1991), Rowe et al. (1994) and by Rowe and Hill (1996). Briefly, workshops were organised at which experienced teachers identified indicators of student progress and possible contexts and methods for the assessment of learning outcomes. These indicators were refined by curriculum specialists and the final set was the result of extensive field testing and statistical analysis using item response modelling. Calibration of the indicators was undertaken to enable them to be organised into levels or bands of roughly equal difficulty as indicated by the logit value range of the indicators on growth continua or strands, of which there are three in English (reading, writing and spoken language) and two in Mathematics (number and space).

In making use of Subject Profiles, teachers are encouraged to use a wide range of both formal and informal assessment methods in arriving at professional judgements regarding students' achievement levels, including direct observations of student performance, written assignments, class-based and standardized tests, specially-constructed assessment tasks, and so on. In other words, the Victorian Profiles function as a framework for teacher assessment and do not in themselves constitute an assessment method. They do, however, provide a means whereby teacher assessments of student performance on the curriculum as taught in schools can be reported using criteria that are consistent across classrooms, schools and the school system.

The manuals accompanying the Victorian Profiles contain extensive advice on their use, including samples of student work to assist teachers in making consistent judgements about standards. In addition, in the early years of their implementation, extensive in-service training was provided, particularly in the use of the English Profiles in primary schools (see Griffin, 1990; Griffin & Nix, 1991; Rowe, 1992a,b; Rowe & Hill, 1996).

Within the context of the VQSP, teachers were required to rate each student's level of achievement with reference to the indicators for each of the nine Bands (A - I) of the reading, writing and spoken language strands of the English profiles, and for each of the twelve Levels (1 - 12) of the number and space strands of the Mathematics profiles. A '3' was recorded if all of the behaviours associated with a given band/level were consistently displayed by the student, '2' if most of the behaviours were present, '1' if some of the behaviours were beginning to be developed, and '0' if none of the behaviours had yet been observed. These ratings for each band/level were then added together to give a total score out of 27 for each of the English profile strands, or 36 for both Mathematics profile strands. A weighted composite score was then constructed for English achievement using the reading, writing and spoken language scores, and for Mathematics achievement using the number and space profile scores.
A key assumption underlying the Victorian Profiles is that they form a cumulative scale similar to that described by Guttman (1944). Using the Guttman method of scaling, lower bound estimates of 'true reliability' for each strand of the English and Mathematics Profiles were computed for large samples of students at each year level (Preparatory to Year 11) and are summarised in Table 1 below. The results indicate that the profiles do indeed function as cumulative scales or growth continua and that teachers are consistent in their use of the scales. The reliability estimates for the earlier years of schooling are not as high as those for the later years on account of the restricted range in the achievement levels of students. In addition, with the exception of the Preparatory Grade, estimates for the two Mathematics strands are somewhat higher than for the three English strands.

Table 1. Guttman Reliability Estimates for the Victorian English and Mathematics Profile Strands, by Grade/Year Level

<table>
<thead>
<tr>
<th>Year Level</th>
<th>English Profile Strands</th>
<th>Maths Profile Strands</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reading</td>
<td>Writing</td>
</tr>
<tr>
<td>Prep (K) (n = 2281)</td>
<td>.791</td>
<td>.740</td>
</tr>
<tr>
<td>Year 1 (n = 1965)</td>
<td>.754</td>
<td>.749</td>
</tr>
<tr>
<td>Year 2 (n = 2188)</td>
<td>.769</td>
<td>.747</td>
</tr>
<tr>
<td>Year 3 (n = 1876)</td>
<td>.800</td>
<td>.786</td>
</tr>
<tr>
<td>Year 4 (n = 2209)</td>
<td>.843</td>
<td>.838</td>
</tr>
<tr>
<td>Year 5 (n = 2015)</td>
<td>.831</td>
<td>.829</td>
</tr>
<tr>
<td>Year 6 (n = 5062)</td>
<td>.845</td>
<td>.833</td>
</tr>
<tr>
<td>Year 7 (n = 3661)</td>
<td>.902</td>
<td>.904</td>
</tr>
<tr>
<td>Year 8 (n = 2630)</td>
<td>.876</td>
<td>.865</td>
</tr>
<tr>
<td>Year 9 (n = 3570)</td>
<td>.926</td>
<td>.927</td>
</tr>
<tr>
<td>Year 10 (n = 2687)</td>
<td>.876</td>
<td>.859</td>
</tr>
<tr>
<td>Year 11 (n = 730)</td>
<td>.898</td>
<td>.922</td>
</tr>
</tbody>
</table>
Further evidence regarding the reliability of teacher assessments using the Victorian Profiles is available in the form of test/re-test reliabilities and inter-rater reliability estimates for the Reading strand of the English Profiles as reported by Rowe (1991a) and summarised in Table 2. The test/re-test reliability estimates presented are Pearson product-moment correlations between teacher assessments of the same students made four months apart in May and September. They indicate that teachers assess their students consistently when asked to provide a repeat assessment.

Table 2. Test/Re-Test and Inter-Rater Reliability Estimates* for the Victorian Reading Profile Strand, by Five Grade/Year Levels

<table>
<thead>
<tr>
<th>Year Level</th>
<th>Test/Re-Test Reliability</th>
<th>Inter-Rater Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 1</td>
<td>0.892</td>
<td>0.855</td>
</tr>
<tr>
<td>Year 3</td>
<td>0.908</td>
<td>0.893</td>
</tr>
<tr>
<td>Year 5</td>
<td>0.911</td>
<td>0.871</td>
</tr>
<tr>
<td>Year 7</td>
<td>0.927</td>
<td>0.832</td>
</tr>
<tr>
<td>Year 9</td>
<td>0.929</td>
<td>0.848</td>
</tr>
</tbody>
</table>

* Pearson product-moment correlations

While the evidence referred to above indicates that teachers are internally consistent in assessing their students when using the Victorian Profiles, it does not address the issue of the extent to which there is consistency across teachers in their assessments of students. Limited evidence regarding inter-rater reliability is presented in the third column of Table 2. The estimates reported are Pearson product-moment correlations between the ratings of two or more teachers who rated the same student. These data were derived from naturally occurring instances (mostly team-teaching situations) in which two or more teachers in the same school were able to provide an assessment of the same student. The results indicate a satisfactory level of inter-rater reliability among teachers using the Victorian English Profiles, reading band, although it must be pointed out that the data are based on observations of teachers who were not responding completely independently, since as colleagues at the same school they were most likely to have developed a common understanding of the performance of their students. This limitation highlights one of the major difficulties in the use of teacher assessments, namely that of obtaining accurate estimates of inter-rater reliability in situations where typically only one teacher observes a student’s performance.

To summarise, the evidence suggests that when using the Victorian Profiles, teachers are highly consistent in their assessment of students and are also able to achieve a satisfactory level of inter-rater reliability (although the evidence is only partial on this point). It is also relevant to note that among teaching staff participating in the VQSP there is agreement that assessments based on the Victorian Profiles have a high degree of validity in terms of measuring students’ levels of achievement in English and mathematics.
as taught in their schools, with such views being most frequently expressed for primary school English and least frequently for secondary school mathematics.

**Student Ability**

Notwithstanding the decision to use teacher assessments as the primary measure of student achievement, it was considered important to obtain a general measure of student ability using conventional standardized tests and to assist teachers in recording students' achievements on the Subject Profiles. This measure consisted of both a verbal and a quantitative component. For the measure of verbal ability, students at and above Grade 3 in 1993 and 1994 were administered one of the TORCH Tests of Reading Comprehension (Mossenson, Hill & Masters, 1987). The TORCH tests are a set of 14 untimed reading tests for use with students in Year Levels 3 to 11 that assess the extent to which readers are able to obtain meaning from text. The tests have been vertically calibrated to provide measures of reading ability on a common scale that ranges from zero to 100.

For the measure of quantitative ability, 20 age/stage appropriate items were selected from the ACER Test of Basic Skills (ACER, 1991) for primary students and a set of 40 items was selected from the ACER Mathematics Profile Series: Review Test (ACER, 1983) for secondary students. For logistical reasons, the tests of quantitative ability were only administered during the second phase of the project (1993) to primary students who were then in Grades 3 and 5, and to secondary students in Years 8 and 10.

**Student home background characteristics**

To obtain measures of intake characteristics of students, the Grade Level, and Gender of each student were recorded (coded '0' for males ans '1' for females). In addition, students residing in areas/towns with a population of less than 25,000 were identified as Rural (coded '1' for rural and '0' for non-rural), and whether or not a student was identified as an Aboriginal or Torres Strait Islander was coded '1' and '0' respectively.

For the 1992 and 1993 data collections, a variable called Socio-Educational Level was measured for each student's family was computed as a weighted composite comprising: the higher of mother's or father's occupational status using the Australian Bureau of Statistics 8-point classification scale (Castles, 1986); the higher of mother's or father's number of years of education, and whether the student was in receipt of the Educational Maintenance Allowance (EMA) – a welfare payment used as an indicator of low income disadvantage (coded '1' for non-EMA and '0' for EMA). Non-English Speaking Background (NESB) was measured as a weighted composite of each student's, mother's and/or father's country of birth, and whether English was the main language spoken at home (coded '1'; '0' otherwise).

For the 1994 data collection, a revised variable called Socio-economic Status was measured as a weighted composite comprising: the higher of mother's or father's occupational status as indicated on the ordinal scale of Item 10 of the Student Background Record (see Appendix “C”), and whether the student was in receipt of the Educational Maintenance Allowance (EMA, coded '1' for non-EMA and 0 for EMA). Non-English Speaking Background (NESB) was indicated as whether or not English was the main language spoken at home (coded '1' and '0' respectively).
Disability/Impairment and 'Critical events'

In the second and third years of the study (1993 and 1994), indicators of student disability/impairment were obtained, including whether or not a student with physical, social or emotional 'disabilities' was integrated into a class (coded ‘1’ for ‘disability’ and ‘0’ otherwise). Consistent with the work of Wallerstein (1991) and Hill et al.'s (1993a) note that student learning outcomes can be seriously "...influenced by critical events, including illness, absence from school, family and personal trauma events including family breakup, loss of employment, death or illness of close relatives, and so on" (p. 32), indicators of critical events were also obtained. Students were assigned a code of ‘1’ (‘0’ otherwise) if one or more of the following ‘events’ had been experienced during that school year: extended illness and/or absence from school, death of a close friend or relative, serious family problems including parental separation or divorce, personal psychological trauma, and loss of job or income within the family.

Student behaviour

Over the three data-collection phases of the study, students' externalizing behaviours in the classroom were measured using teachers' ratings on the 16 bipolar items of the Parent/Health Professional form of the Rowe Behavioural Rating Inventories (RBRI - see Rowe & Rowe, 1989, 1993, 1995a,b). Following the item format advocated and used by Kysel, Varlaam, Stoll and Sammons (1983), response items for the RBRI instruments were developed to allow for both positive and negative assessments of student/child externalizing behaviors of concern to parents and teachers, in three domains: Irritable-Antisocial/Sociable, Inattentive/Attentive and Restless/Settled.

The RBRI instruments were devised specifically for: (1) monitoring the comorbidity of externalizing behaviours and academic under-achievement in longitudinal educational/psychosocial research (ie., Hill et al., 1993a,b; Hill & Rowe, 1996; Rowe, 1991a,b, 1995; Rowe, Hill & Holmes-Smith, 1995; Rowe & Rowe, 1992b, 1994b, 1995), and (2) for use in clinical/epidemiological studies designed to examine the relationship between diet and behaviour (see Rowe, 1988; Rowe & Briggs, 1992, 1993; Rowe & Rowe, 1994). The rationale for using these scales, together with their utility in research applications, has been comprehensively documented in the references cited here.

Using a semantic bipolar format, the item nomenclatures were formulated on the basis of extensive cross-validations of parent and teacher descriptions of in-context behaviors at home and at school. For this project, the items, measured on 5-point ordinal scales, were coded in the positive direction, from negative to positive behaviours (ie., 1-5). In the first phase of the project (1992), parent ratings of their child's behaviour at home were also obtained to allow for comparative analyses of concurrent parent-teacher ratings on these three domains. Scale scores for the three domains were computed as weighted composites of their constituent item indicators.

Student attitudes

Measures of students' attitudes towards schooling and their perceptions of the "quality of school life" were obtained from responses on an inventory adapted from instruments developed by Ainley, Reed and Miller (1986); Ainley, Goldman and Reed (1990). From multiple items (each measured on 4-point ordinal scales), weighted composite scores
were computed for each of five scales: Enjoyment (subsequently referred to as Attitude to School), Social Acceptance, Teacher Responsiveness, Teacher Expectations, and perceived Curriculum Usefulness.

**Homework**

On six items, students indicated the frequency with which they undertake homework in English, mathematics and other subjects – each measured on 5-point ordinal scales ranging from "Every day", "2-3 times per week", "Once per week", "2-3 times per month, and "Never". Again, weighted composites for English Homework and Maths Homework were computed.

**Classroom instruction**

For all three data-collection phases, use was made of two classroom-level variables. The first was for elementary schools only and was a dummy variable coded '1' if the class contained students from two or more Grade Levels and '0' if all of the students in the class were from the same Grade Level. In recent years, the organization of students into what are referred to as Composite Classes has become common in the Victorian context, both in response to unequal numbers of student enrolments within any one Grade level and as a deliberate policy to promote multi-age classroom teaching. The second variable was a measure of recent teacher participation in structured professional development In-Service Programs. At the primary level, these programs were almost entirely related to specific programs such as the Canberra Literacy In-Service Program (CLIP), Early Literacy In-Service Course (ELIC), Early Mathematics In-Service Course (EMIC), Key Group Maths, and Reading Recovery. At the secondary level, the programs attended by teachers were typically more varied in focus and relate generally to the teaching of English and Mathematics. Again, this was a dummy variable coded '1' if the teacher had recently participated in such a program and '0' otherwise.

Further, in 1994, students' responses were obtained on a questionnaire designed to measure their opinions about the quality of teaching and learning they experienced. Drawing, in part, on the theoretical QAIT model of instructional effectiveness elaborated by Slavin (1987, 1994), the questionnaire consisted of sets of items relating to the following nine sub-scales: teacher Empathy, Energy/Enthusiasm, Fairness/Firmness, Helpfulness/Responsiveness, High Expectations, Quality of Instruction, Feedback, Appropriateness of Instruction and Time. Again, each scale was computed as a weighted composite of its constituent indicator items as indicated in Appendix "C".

**Parent responses**

On a Parent Questionnaire administered in 1992 and 1993 (see Appendices "A" and "B"), parent(s)/guardian(s) provided data for measures in six major areas: (1) family socio-educational background factors (see above); (2) estimates of the amount of time students spend on school-related and extra-curricula activities per week; (3) the source and frequency of help given to students with regard to homework; (4) the type and frequency of participation in school activities related to curriculum assistance (Assisting with Class Activities), attendance at both parent-teacher interviews and information evenings (Attendance at Parent-Teacher Interviews, etc.), and school administration (Involvement in Decision-Making), (5) an evaluation of the school in terms of the

Centre for Applied Educational Research, The University of Melbourne
quality of information received on students' educational progress and the perceived extent to which the school meets the educational need of their child (*Parent Perceptions of the School*) and, finally, (6) a rating of their child's externalizing behaviours at home using the RBRI 20-item Parent Form (Rowe & Rowe, 1989, 1993, 1995a,b) – the first 16-items of which are identical to the Teacher/Health Professional Form completed by teachers.

The *Parent Questionnaire* was modified for the 1994 data-collection (see Appendix “C”) to ensure consistency with an instrument developed jointly by the Centre for Applied Educational Research and the Directorate of School Education, Victoria, as part of their *Schools of the Future* program (Victoria, 1993). The questionnaire consisted of three sections: (a) Parent opinions of the school – consisting of 5 sub-scales (quality of teaching and learning; academic and social ethos; general school environment; management and responsiveness; reporting progress), (b) Parental involvement with the school – made up of 4 sub-scales (school administration/policy; interest in child's progress; assistance with classroom activities; attendance at parent-teacher interviews); and (3) Homework – consisting of 2 sub-scales (parental help with homework; homework conditions).

**Teacher affect**

Measures were obtained from responses on a semantic differential instrument consisting of 20, seven-point evaluative scales. This instrument was adapted from the *Professional Self-Perception Questionnaire* (PSPQ), originally developed by Elsworth and Coulter (1977) and modified by Rowe and Sykes (1989). Two scales from the PSPQ were used in this study, namely, *Energy/Enthusiasm* and *Warmth* towards students. These two measures were combined to provide a single measure of *Teacher Affect*.

**Teachers' perceptions of their work environment**

Teachers' perceptions of various aspects of their work environment were measured using items from the *School Organisational Health Questionnaire* (SOHQ) developed by Hart, Conn and Carter (1992). Specific details of the development of the SOHQ and its psychometric properties have been documented by Hart, Carter, Conn, Dingle and Wearing (1993), and aspects of its utility have also been reported by Rowe and Hill (in press). The SOHQ consists of 12 sub-scales, namely: *Curriculum Co-ordination, Decision Making, Discipline Policy, Feedback, Goal Congruence, Leadership Support, Morale, Peer Support, Professional Development, Role Clarity, Student Orientation and Work Demands*.

**Aspects of leadership**

Given the claimed importance of the impact of principal leadership in school effectiveness research (see Hallinger & Leithwood, 1994), a *Leadership Questionnaire* was designed and administered to principals and deputy principals in participating schools during the second and third phases of the project (1993, 1994). This instrument was developed to obtain measures of various aspects of the work undertaken by school leaders as part of their role. These aspects, *inter alia*, included: Leadership Style Values (*transformational, transactional, public choice*) – influenced by Leithwood's notion of the transformational leader (Leithwood, 1992; Leithwood, Begley & Cousins, 1992;
Analytic methods

Descriptive statistics

Descriptive statistics (i.e., first-, second- and third-order moments) for all variables of interest were calculated using standard techniques. Where appropriate, these data have been presented in both tabular and graphical forms (e.g., histograms, box-and-whisker plots; see Tukey, 1977).

Construct measurement and reliability

To ensure that measurement error problems in observed indicator variables were minimised for fitting explanatory models to the data (see below), a specific approach to computing composite variables was used that has been developed for specific applications in explanatory research work of the present kind (see Holmes-Smith & Rowe, 1994). This approach is described below.

Most explanatory models in applied educational and psychosocial research are formulated in terms of hypothetical constructs (or latent variables) that are not directly measurable or observable. As a means of data reduction, it is commonplace to compute latent or composite variables such as ability, motivation, socio-economic status, attitudes, and so on, from several observed indicators, each measured on dichotomous or Likert-type ordinal scales. Traditionally, and mostly using Pearson product-moment inter-item correlation estimates, such composites have been computed as factor scores.

When observed variables are non-normal and non-continuous (e.g., ordinal, polytomous categories), the use of product-moment correlations is inappropriate (Jöreskog, 1990), yielding large negative biases in their estimates (Carroll, 1961). It is important to note that in general, SEM techniques assume that observed variables are quantitative variables measured, at least approximately, on an interval scale, and whose distributions are approximately multi-normal. In most educational and psychosocial research applications, however, the observed variables are typically non-normal and/or of mixed scale types: categorical, ordinal (Likert-type rating scales) and continuous. Under such circumstances, the use of ordinary product-moment correlations is not appropriate (see Healy & Goldstein, 1976). Instead, tetrachoric (categorical with categorical) polychoric (ordinal with ordinal) and polyserial correlations (continuous with ordinal) should be computed, and the correct asymptotic covariance matrix of such correlations should be analysed by the method of Weighted Least Squares (WLS), using PRELIS (Jöreskog & Sörbom, 1988, 1993b). Failure to do otherwise can lead to gross errors in correlation estimates, distorted structural equation parameter estimates, and incorrect goodness-of-fit measures and standard errors (Huba & Harlow, 1987; Jöreskog & Sörbom, 1989a, 1989b, 1993a, 1993b).

From Jöreskog (1994, p. 383), the special features of ordinal variables are worth noting:

Observations on an ordinal variable are assumed to represent responses to a set of ordered categories, such as a five-category Likert scale. It is only assumed that a person who responds in one category has more of a characteristic than a person who responds in a lower category. Ordinal variables are not continuous variables and should not be treated as if they are. Ordinal variables do not have origins or
(derived from exploratory factor analysis or principal-component analysis), or as simple, unit-weighted, additive indices of the raw indicator scores, regardless of either the measurement and distributional properties of the constituent variables, or their relative contribution to the composites. Typically, the derived indices for the composite variables are then treated as continuous variables in omnibus general linear model techniques, which assume that such indices are measured without error.

This approach leads to at least two problems when trying to model relationships among composite scales. First, the unit-weight addition of observed raw scores on indicators to compute scale scores ignores the possibility that some indicators may contribute more to the measurement of the underlying latent trait than others. Second, the unit-weight addition of raw indicator scores may invalidate the composite scale if one or more of the indicators 'measure' a latent trait other than the one under consideration.

In recent years, these problems have been reduced somewhat by the use of structural equation models (see Bentler, 1980; Bollen, 1989; Jöreskog & Sörbom, 1979; McDonald, 1978; Muthén, 1994). The measurement part of these models allow for unequal contributions of indicator variables towards the measurement of latent variables and the models will fit only when the indicator variables associated with any one latent variable are valid indicators of that construct. However, when the number of indicator variables becomes large (as in the present case), parameter estimation and model fit statistics are unstable unless the sample size is also large.

In the VQSP, extensive use was made of one-factor, congeneric measurement models to maximise the reliability of composite scales and latent variables for subsequent use in fitting explanatory models to the data. Unlike traditional unit-weighted methods for computing composites, the use of factor score regression weights obtained from such one-factor models minimises measurement error in the items contributing to each scale, thus increasing the reliability (and validity) of the computed scale scores. For explanatory research applications, the use of maximally reliable composite scores is crucial in fitting both single- and multi-level regression models (Bryk & Raudenbush, 1992; Goldstein, 1995), as well as in fitting structural equation models (Bollen & Scott Long, 1993; Jöreskog & Sörbom, 1993c).

units of measurements. Means, variances, and covariances of ordinal variables have no meaning (my emphasis).

It is common practice to treat scores 1, 2, 3, 4, representing the ordered categories of an ordinal variable as numbers on an interval scale and use a covariance matrix computed in the usual way to estimate a structural equation model. What is so bad with this is not so much that the distribution is non-normal; more importantly the distribution is not continuous: there are only four distinct values in the distribution. The use ordinal variables in structural equation models requires other techniques than those which are used for continuous variables.

Cuttance (1987, pp. 245-250), Jöreskog (1990, 1994), Jöreskog and Aish (1993), and Muthén (1993) provide detailed discussions on the treatment of categorical, ordinal and non-normal variables in structural equation models with ordinal variables. In developing the WLS method available in LISREL 7 and LISREL 8, to assist in minimising problems with non-normally distributed variables, Jöreskog and Sörbom provide a method for obtaining an appropriate weight matrix, correct parameter estimates, standard errors and a fit statistic. "The weight matrix required for such an analysis is the inverse of the estimated asymptotic covariance matrix W of the polychoric and polyserial correlations" (Jöreskog & Sörbom, 1993b, p. 45).

Centre for Applied Educational Research, The University of Melbourne
Using either LISREL 7 or LISREL 8 (Jöreskog & Sörbom, 1989a, 1993a,c) under a weighted least squares method of estimation and a listwise method for deleting missing data, the one-factor models were fitted to the constituent ordinal-scaled, item data, based on a scaled covariance matrix (and its asymptotic estimates) of the polychoric correlations computed by PRELIS 1 or PRELIS 2 (Jöreskog & Sörbom, 1988, 1993b) (see Note 5). Composite scores computed by this method are single indices of their component items, each of which is weighted for its relative contribution to the composite.

The one-factor, congeneric, measurement model (i.e., LISREL Sub-Model 1) may be illustrated diagrammatically for the Morale scale from the Teacher Questionnaire, for example, as shown in Figure 1.1, where its constituent items (given below) require responses on Likert-type, 5-point ordinal scales:

<table>
<thead>
<tr>
<th>Item</th>
<th>Statement</th>
<th>Strongly disagree</th>
<th>Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(x_1; i25) There is a good team spirit in this school</td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>(x_2; i32) There is a lot of energy in this school</td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>(x_3; i40) The morale in this school is high</td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>(x_4; i48) Teachers go about their work with enthusiasm</td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>(x_5; i54) Teachers take pride in this school</td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
</tbody>
</table>

![Diagram of the one-factor, congeneric measurement model for Morale]

Figure 1.1 One factor, congeneric measurement model for Morale

In matrix format, Figure 1.1 shows the regression of \( x_i \) on \( \xi_1 \) where the elements \( \lambda_{xi} \) are the partial regression coefficients of \( \xi_1 \) in the regression of \( x_i \) on \( \xi_1 \), namely:

\[
\begin{bmatrix}
  x_1 \\
  x_2 \\
  x_3 \\
  x_4 \\
  x_5 \\
\end{bmatrix}
= \begin{bmatrix}
  \lambda_{11} \\
  \lambda_{21} \\
  \lambda_{31} \\
  \lambda_{41} \\
  \lambda_{51} \\
\end{bmatrix}
\begin{bmatrix}
  \xi_1 \\
\end{bmatrix}
+ \begin{bmatrix}
  \delta_1 \\
  \delta_2 \\
  \delta_3 \\
  \delta_4 \\
  \delta_5 \\
\end{bmatrix}
\]

or,

\[ x_i = \lambda_{xi} \xi_1 + \delta_i \quad [1] \]

The assumed model implies that the covariance matrix of the observed indicators (\( x_i \)) is of the form:

Centre for Applied Educational Research, The University of Melbourne
\[ \Sigma = \lambda_1 \lambda_1' + \Theta_\delta \]  

where \( \Theta_\delta \) is a diagonal matrix with elements \( \theta_{\delta i} \) indicating the variances of \( \delta_i \) (i = 1, 2, 3, 4, 5). From the parameters of equation [2] the reliability \( (r_c) \) of a composite \( (\xi_c) \) is given as

\[ r_c = \frac{w_c' (\Sigma - \hat{\Theta}_\delta) w_c}{w_c' \Sigma w_c}, \]  

where \( w_c \) is a vector of factor score (FS) regression weights that maximise the reliability of the composite. For specific details of these well-established but all too rarely used procedures, see Alwin and Jackson (1980), Brown (1989), Fleishman and Benson (1987), Jöreskog (1971), Munck (1979), Werts, Rock, Linn and Jöreskog (1978). Further details including the rationale for this approach to computing composite variables and their reliabilities have more recently been outlined and demonstrated by Hill et al. (1993a); Holmes-Smith and Rowe (1994); Rowe (1995a,b, 1996); Rowe, Hill and Holmes-Smith (1995), and by Rowe and Hill (in press).

The results of fitting the above congeneric model to the data are as follows:

![Figure 1.2 Solution to one-factor model for Morale](image)

Model 'goodness-of-fit' indices:  
\( \chi^2 (4) = 5.99, p = .200 \)  
GFI = 1.000  
AGFI = .999  
RMR = .007

**Total Coefficient of Determination** \( (R^2) = 0.902 \)

Here, \( R^2 \) is equivalent to the composite scale reliability coefficient \( (r_c) \) calculated from the maximally weighted factor score regression coefficients obtained from fitting the one-factor congeneric measurement model to the constituent scale items, and estimated by equation [3], ie:

\[ r_c = \frac{w_c' (\Sigma - \hat{\Theta}_\delta) w_c}{w_c' \Sigma w_c}, \]
where $\Sigma$ is the estimated variance-covariance matrix of the indicators; $\Theta_6$ is estimated variance-covariance matrix among their measurement errors, and $w$ is the vector of the related factor score regression coefficients (FS). $^6$

$$[0.140, 0.206, 0.205, 0.330, 0.224],$$

for $i25$, $i32$, $i40$, $i48$ and $i54$, respectively.

A proportionally weighted scale score for the composite variable Morale that takes into account the individual and joint measurement error of the indicators $i25$, $i32$, $i40$, $i48$ and $i54$, can now be computed as a continuous variable for each case as follows (eg., using SPSS format nomenclature):

```
compute MORALE = (i25*0.127)+(i32*0.186)+(i40*0.185)+(i48*0.299)+(i54,*0.203)
```

where $i25$, $i32$, $i40$, $i48$ and $i54$ are the raw score ratings made by each respondent on the five indicator items respectively. This process ensures that the estimation of the scale/composite variable Morale (adjusted for measurement error) is proportionally weighted by the actual contribution made by each indicator. Note that these proportionally weighted FS regression coefficients add to 1; hence the scale/composite score will range from a minimum of 1 to a maximum of 5. This means that the composite variable Morale ($\xi_{21}$), and all other scales computed similarly, have the advantage of all being ‘measured’ in the same metric. The computed scale/composite score for Morale (and similarly in the case for each construct) may now be used in exploratory analyses, namely, in fitting explanatory models, including structural equation models and multilevel models.

**Fitting explanatory multilevel models**

Reference has already been made to the paucity of school effectiveness studies that have accounted for the inherent multilevel organizational structure of schooling (see pp. 3-5). In fact, the failure of many studies to account for the typical hierarchical or multilevel structure of the data has been a prominent theme in methodological criticisms of epidemiological, economic, educational and psychosocial research over the past twenty years. $^7$ This is especially the case for the analysis of data in school effectiveness research

For a one-factor congeneric measurement model, the factor score regression coefficients (FS) represent the estimated bivariate regression of the factor ($\xi$) on all the observed indicator variables, given by: $FS = \Lambda \Sigma^{-1}$, where $\Lambda$ is the estimated factor pattern matrix and $\Sigma$ is the estimated covariance matrix of the observed indicators (see Jöreskog & Sörbom, 1989b, p. 93; Lawley & Maxwell, 1971, p. 109). Factor score estimates ($\xi$) may be computed for any individual $i$ with observed scores $x_i$, using the simple product: $\xi = FS x_i$.

Compared with their counterparts in economics, education, epidemiology, sociology, and the biological sciences, researchers in social and organizational psychology, however, have been slow to acknowledge the implications for their own research applications of the now large literature on random coefficient models where population units are hierarchically clustered. For example, recent applications in economics (Langford, Bateman & Langford, 1996), epidemiology (Langford & Bentham, 1996) and sociology (Lee, Dedrick & Smith, 1991; Willms & Paterson, 1995), are worth noting. The books by Bock (1989), Raudenbush and Willms (1991) contain a range of applications, and those by Bryk & Raudenbush (1992), Goldstein (1987, 1995) and Longford (1993) provide details of the related estimation theory with annotated examples. A special issue of the Journal of Educational and Behavioral Statistics edited by Ita Kreft (1995) focuses on issues related to the application of multilevel models in psychosocial research.

Centre for Applied Educational Research, The University of Melbourne
(Bock, 1989; Bosker, Creemers & Scheerens, 1994; Bryk & Raudenbush, 1989, Cuttance, 1992; Goldstein, 1987; Hill & Rowe, 1996; Jesson & Gray, 1991; Kang, 1992; Raudenbush, 1989, 1995; Raudenbush & Willms, 1991; Rowe, 1990a, Rowe & Hill, 1994; Rowe et al., 1995; Scheerens, 1993). Units with a multilevel structure (i.e., students within classes within schools) give rise to observations that fail to satisfy the assumption of independence required for fitting single-level linear models (Bryk & Raudenbush, 1992; Goldstein, 1987, 1995; Longford, 1993).

It has long been recognised that the existence of such clustering poses special problems related to levels of analysis that lead to several long-standing and troublesome obstacles to statistical conclusion validity (see Robinson, 1950; Cronbach & Webb, 1975). Such obstacles include: aggregation bias, undetected heterogeneity of regression, misestimated parameter estimates and their standard errors, and related problems of model misspecification due to lack of independence between measurements at different levels (Aitkin & Longford, 1986; Burstein, 1980, 1988; DeLeeuw & Kreft, 1986; Goldstein, 1986; Mason, Wong & Entwistle, 1984; Raudenbush & Bryk, 1986, Rousseau, 1985). Moreover, failure to account for the essential hierarchical structure of the data typical of most psychosocial research, invariably leads to an increased probability of committing Type I errors (i.e., falsely rejecting the null hypothesis), with important ramifications for the substantive interpretation of findings and their related policy implications (Aitkin, Anderson & Hinde, 1981; Rowe, 1992). Unfortunately, such errors are not trivial and occur all too frequently in psychosocial research. Moreover, there is often important information about variation within and between clusters that require the application of new kinds of statistical models for their analysis.

During the past ten years, considerable progress has been made towards the solution of these analytic problems. Presently, there are at least five statistical software packages available for analysing multilevel data, including: (1) GENMOD (Hermalin & Anderson, 1989; Wong & Mason, 1985); (2) HLM-3L (Bryk, Raudenbush & Congdon, 1992); (3) ML3/MLn (Prosser, Rasbash & Goldstein, 1993; Rasbash & Woodhouse, 1995, 1996), (4) VARCL (Longford, 1987, 1988), and (5) BMDP-5V (Schluchter, 1988). Comparative reviews of these packages are provided by Kreft, de Leeuw and Kim (1990), Kreft, de Leeuw and van der Leeden (1994) and by Yang (1992). Technically, there are few limitations on the kinds of models that can now be fitted to account for the full hierarchical structure of data typically obtained in psychosocial and epidemiological research, especially since the release of MLn (Rasbash & Woodhouse, 1995, 1996) with its capacity to analyse any number of nested levels, including cross-classifications.8

8 The preference for MLn over the alternative statistical packages cited here is due principally to its extreme efficiency and interactive flexibility. Moreover, the on-going software development of MLn, together with the proactive training and user-support facilities provided by the Multilevel Models Project group at the University of London’s Institute of Education, are major advantages for researchers. On the basis of considerable experience the following review comments of Kreft, de Leeuw and van der Leeden (1994) are strongly endorsed:

Of the multilevel programs, ML3 (MLn) allows users the most freedom to choose input and even to make adjustments during the run (p. 329), and

We think ML3 (MLn) is the most appropriate program for serious users and certainly for people doing research in multilevel analysis (p. 344).
To estimate the proportion of variance in the response and explanatory variables of interest due to class/teacher contextual effects, multilevel models were fitted to the maximally-weighted composite variables, at each time period, as follows:

(a) Three-level variance-components models (hereinafter referred to as the "null models"), to estimate the variance due to the group effects of students (level 1) within classes/teachers (level 2) within schools (level 3), for each of the response variables of interest,

(b) Three-level, "intake-adjusted" regression models, with the student intake characteristics and mediating variables fitted as fixed, level 1 explanatory variables, and the contextual variables fitted as fixed, level 2 or level 3 explanatory variables.

To examine the stability of the measures for student achievement (say), as well as the effects of contextual factors, two types of multilevel, longitudinal models have been fitted to the data: (1) fixed occasion, repeated measures models, and (2) polynomial growth curve models. (For specific details of these models, see Bryk & Raudenbush, 1993; Goldstein, 1989, 1995).

For convenience here, equations for the null models (a) are illustrated. Following Rasbash & Woodhouse (1995), this model can be written in three parts. First, for the \( i \)th student in class/teacher \( j \) within school \( k \), scores on each of the externalizing behavioural domains, or for Literacy and Numeracy achievement \( Y_{ijk} \), can be modelled as

\[ Y_{ijk} = \beta_{0jk}X_0 + e_{ijk}. \]  

Here, the slope is constant (0 - zero) but the intercept is random, varying across classes/teachers and schools. The \( X_0 \) term in equation [4] is a column vector of unities representing the constant slope for classes/teachers and schools, and \( e_{ijk} \) is a random residual term representing the contribution to the response variable \( Y_{ijk} \) of the \( i \)th student in the \( j \)th class/teacher within the \( k \)th school.

Second, the intercept for class/teacher \( (\beta_{0jk}) \) can be expressed as a linear function of the average intercept for school \( k \) \( (\pi_{00k}) \) and a class/teacher-level random term \( (u_{0jk}) \):

\[ \beta_{0jk} = \pi_{00k} + u_{0jk}. \]  

Third, the average intercept for school \( k \) can be modelled in terms of an overall school average intercept \( (\gamma_{000}) \), and a school-level random term \( (\nu_{00k}) \):

\[ \pi_{00k} = \gamma_{000} + \nu_{00k}. \]  

By combining equations [4], [5] and [6] a single equation version of the model can be written as follows:

\[ Y_{ijk} = \gamma_{000} + (\nu_{00k} + u_{0jk} + e_{ijk}). \]  

where \( \gamma_{000} \) is the fixed part of the model and the three random terms are bracketed.
From equation [7], given the fixed part of the model ($\gamma_{000}$ - the grand mean of $Y_{ijk}$), the random parameters that are estimated for this model are the variances of the residual terms in brackets, namely:

- the between-school variance estimate of the residual term $v_{00k}$ (i.e., $\sigma_v^2$);
- the between-classes/teachers variance of the residual term $u_{0jk}$ (i.e., $\sigma_u^2$), and;
- the between-students, within classes/teachers and schools variance of the residual term $e_{ijk}$ (i.e., $\sigma_e^2$).

The total variance due to random effects ($\sigma_T^2 = \sigma_v^2 + \sigma_u^2 + \sigma_e^2$) may then be partitioned into that due to school, class/teacher and student effects as follows:

Proportion of variance due to school effects = $\sigma_v^2/\sigma_T^2$; class/teacher effects = $\sigma_u^2/\sigma_T^2$; and student effects = $\sigma_e^2/\sigma_T^2$.

Under an iterative generalised least squares method of estimation (see Goldstein, 1986), models described by equation [7] - and elaborations of it - were fitted to the data using \textit{MLn} (Rasbash & Woodhouse, 1995, 1996).

To estimate the effect magnitude of student-level mediating and explanatory variables, as well as the effects of contextual variables at the class/teacher-level, multivariate multilevel models were be fitted. Specifications of these models are straight-forward extensions of the variance-components models given above. Moreover, the fixed occasion, repeated measures models, and polynomial growth curve models constitute simple extensions of these models again (see Bryk and Raudenbush, 1992; Rasbash & Woodhouse, 1995; Woodhouse, 1996).

\textit{Fitting explanatory, multilevel, structural equation models}

The theory and application of single-level structural equation modelling is well known and used widely in the psychosocial sciences (see Anderson, 1987; Bentler, 1980; Bollen, 1989; Cuttance & Ecob, 1987; Hayduk, 1987; Joreskog & Sorbom, 1979, 1989, 1993c; Marsh, 1994; McDonald, 1985; Rowe, 1991, 1995; Scott Long, 1983). Despite their limitations as single-level models (see below), there are two major advantages in fitting \textit{unconditional} structural equation models (SEM's). First, they provide a means of estimating jointly the magnitude of direct, indirect and interdependent effects among variables. Second, they can account for measurement error in both the observed and latent variables. Nevertheless, as already indicated, when the data to be modelled are hierarchically structured, single-level methods of fitting SEM's are inefficient and inappropriate.

The crucial drawback of fitting \textit{single-level} SEM's to multilevel data is that they do not allow for comparison of relationships across levels. Moreover, such models typically lack power due to the smaller number of observations in the higher-level units. There are, however, ways of minimising these difficulties using multivariate, multilevel models and structural equation models. Related procedures have been outlined and suggested by several authors (see Goldstein, 1995, ch. 4; Goldstein & McDonald, 1988; Goldstein &
Woodhouse, 1996; McDonald & Goldstein, 1989; Muthén, 1994; Raudenbush, 1995; Rowe & Hill, in press). For the present data, an initial approach was to partition the variances and covariances among the five composites into separate school-, teacher- and student-level variance-covariance matrices using multivariate multilevel models, as preludes to fitting explanatory, three-level, structural equation models.

For fitting structural equation models to the data, the LISREL method for Sub-model 3b (Jöreskog & Sörbom, 1989, pp.189-190) was used. This general model contains only y (observed) and η (latent or composite) variables. In fitting such models it is hypothesised that all latent variables (manifest or composite) essentially form part of an interdependent, endogenous system.

The structural relationships among the latent (composite) variables (η) are given by

$$\eta = B\eta + \zeta$$

where $B$ is the matrix of effect relationships among the endogenous constructs (η) and $\zeta$ is the vector of prediction residuals in the structural equations. The major interest in the structural equation models fitted in the study concerned the magnitude estimation of the direct, indirect and total effects for the parameters of $B$.

The measurement model for the observed $y$ variables is given by

$$y = A_y(I - B)^{-1}\zeta + \varepsilon$$

where $A_y$ is a matrix of factor coefficient loadings of $y$ on $\eta$, and $\varepsilon$ is a vector of measurement errors for the $y$ variables. The covariance matrix of $y$ is

$$\Sigma = A_y(I - B)^{-1}\Psi(I - B')^{-1}A_y' + \Theta_{\varepsilon}$$

where $\Psi$ is the covariances among the $\zeta$'s and $\Theta_{\varepsilon}$ is the covariances among the $\varepsilon$'s.

There are several advantages of using this submodel rather than the full LISREL model. First, it may be preferred because it has only four parameter matrices to be estimated, namely, $A_y$, $B$, $\Psi$, and $\Theta_{\varepsilon}$; and provides for correlations among all $\varepsilon$'s. Of greater importance, there are strong substantive grounds for treating both observed ($y$) and latent variables (η) as endogenous. Influenced by recent developments in Chaos theory, there is a growing body of opinion from within modern systems theory (see Simon, 1993) suggesting that all elements within psychosocial 'systems' are endogenous, including a person's gender, socio-economic status, and so on. In fact, psychosocial theorists and researchers are finding increasing difficulty in justifying the nomination of certain variables as exogenous. Besides, once either an observed or latent variable has been specified as exogenous ($x$ and $\xi$ respectively), one cannot estimate the regression effects among them, or estimate path coefficients from endogenous (η) to exogenous (ξ) constructs – by definition! See Rowe and Rowe (1992b) for a substantive research application in the use of such models.

For each of the structural equation models fitted, indices of the estimates for the fixed parameters ($\lambda_c$) of the composite variables were calculated from $\lambda_c = \sigma_x \sqrt{f_c}$, (from
equation 3) and the fixed parameter estimates for the composites' unique (error) variances ($\Theta_c$) were derived from $\Theta_c = \sigma_x^2 (1 - r_c)$.

**Summary of Results**

Results from the VQSP have provided powerful insights into schooling that are particularly illuminating, as well as providing pointers for school improvement and the resolution of major policy issues. Since detailed findings and results of the study are available elsewhere, they are not presented here. Rather, this report provides a summary of the key findings, together with illustrative graphical representations where appropriate. (For specific details of the findings, see Hill et al., 1993a, 1994, 1995; Hill & Rowe, 1996; Rowe & Hill, 1996, in press; Rowe et al., 1993, 1994, 1995). The following key findings have been identified as being of special interest and relevance.

1. **Subject Profiles provide an effective framework for monitoring and reporting students' achievements and progress**

Mention has already been made to the rationale for using Subject Profiles in the VQSP to measure students' achievements/progress in Literacy and in mathematics (see pp. 14-18). An important conclusion drawn from the project over the three years of its data-gathering duration is that in the context of a low-stakes research study, the Victorian Profiles function as effective frameworks for monitoring student progress over Year Prep. (K) to Year 11. In addition to providing a broad-based, 'authentic' and reliable approach to the assessment of student learning outcomes, Profile assessments were sensitive to growth and change over the years of schooling.

To illustrate this, Figures 1.1 to 1.3 summarise both the cross-sectional and longitudinal data for teacher-rated achievement levels of students in each of Years Prep. to 11 on the **reading, writing and spoken language** strands of the **English Profiles** in the form of 'box-and-whisker' plots (Tukey, 1977). The box-plots are used to describe the 'shape' of the distributions for each Year Level. Similar data from teachers' ratings of students' established and developing achievements on the **number and space** strands of the **Mathematics Profiles** are presented in Figures 2.1 and 2.2.

The shaded boxes in Figures 1.1 to 2.2 describe the range of achievement of the 'middle' 50 per cent of students at those Year levels. The top of each box indicates the level of students achieving at the 75th percentile, the bottom of the box shows the 25th percentile and the asterisk indicates the 50th percentile, or median value. The top and bottom 'whiskers' show the 90th and 10th percentile levels of achievement respectively. Approximate 'lines of best fit' have been drawn on each graph for the 10th, 25th, 50th, 75th and 90th percentile values.

The distributions shown in Figures 1.1 to 1.3 for the **reading, writing and spoken language** strands indicate a period of rapid growth in students' achievements during the first few years of schooling, coinciding with the period during which pupils acquire basic skills, and thereafter show a consistent rate of growth to Year 9. It is noticeable, however, that the range of achievement increases markedly over the years of schooling,
with more than four band widths separating Year 9 students at the 10th and 90th percentiles. Of particular concern is the flattening out of the 'growth trajectory' at the 10th percentile, indicating a trend of less than one 'band width' of growth from Year 4 to Year 9. Note also, the minimal growth between Years 9 and 10.

Figure 1.1 Box plots showing distributions for students' progress on the Victorian English Profiles - Reading Strand, by Grade/Year level and cohort.
Figures 1.1 to 1.3 also provide evidence of a discontinuity between primary and secondary schooling for Literacy achievement, with a 'dip' in the rate of progress of students in the first year of secondary school (Year 7). This pattern has been observed in several studies using common measures over primary and secondary schooling (e.g. Elly, 1992; Lunberg & Linnakylä, 1993; Purves, 1973). An interesting feature of this pattern, however, is its striking similarity with that shown by paediatric percentile growth-charts for height and weight during the pre-pubertal to early adolescent period of development. In commenting on this phenomenon Rowe (1995, p. 78) notes: "It is possible that what has become known as an 'educational phenomenon' [i.e. 'apparent dips' in literacy
performance during the transition from primary to secondary schooling] may also have developmental psycho-physiological correlates”.

The most significant finding, however, is the fact that the bottom decile of students is making minimal progress in reading, writing and spoken language between Years 4-9.

Figure 1.3 Box plots showing distributions for students’ progress on the Victorian English Profiles - Spoken Language Strand, by Grade/Year level and cohort

Above all, these findings highlight the crucial importance of the early childhood years in establishing a firm foundation for learning and confirm the views of those who advocate placing a high priority on early intervention programs such as Reading Recovery (Clay, 1985) that have the potential to relocate students on a steeper ‘growth trajectory’ (see Centre for Applied Educational Research, The University of Melbourne
Rowe, 1995; Sylva, Hurry & Plewis, 1995). Another conclusion is that the later years of schooling are less effective in overcoming any disadvantage relating to student background characteristics and initial levels of achievement.

The data summarised in Figures 2.1 and 2.2 (below) indicate a steady and consistent rate of growth in students' achievements in number and space from Prep. to Year 11. However, similar to the distributions shown for reading (Figure 1.1), the increasing differentiation in the range of student achievement from primary to secondary schooling indicates that secondary teachers are required to provide and manage teaching and learning activities and strategies for a wider range of student competence than that required of their primary colleagues.

Figure 2.1 Box plots showing distributions for students' progress on the Victorian Mathematics Profiles - Number Strand, by by Grade/Year level and cohort

Centre for Applied Educational Research, The University of Melbourne
Information of the type shown in Figures 1.1 to 2.2 is of interest in its own right by providing a ‘picture’ of student progress within specific curriculum areas. As noted by Hill (1994): “It is also essential to enable Subject Profiles to be used to report student achievement and progress in terms of standards typical of students of the same age or year cohort” (p. 39).

2. **There are major gender differences on all indicators affecting students’ progress**

Consistent with a growing body of research, findings from the VQSP indicated large differences between male and female students on all key factors affecting learning outcomes. In sum, girls in all five cohorts indicated significantly higher levels of achievement and rates of progress than males, and had more positive attitudes towards school and behaviours in the classroom. These findings are illustrated in Figures 3 to 5.
The data presented in Figure 3 are equivalent of those shown in Figure 1.1, but clearly illustrate the superior achievement and progress in reading by girls at all Year levels—especially during the secondary school years. Similar patterns were also evident on the related literacy strands (writing and spoken language), as well as on the mathematics strands of number and space.

Figure 4 presents the comparative data for male and female students' Enjoyment of school. Overall the data indicate that students enjoy being at school, but show a notable decline in Enjoyment between Year 3 and Year 9, particularly among boys.
Figure 4. Box plots showing distributions for students' attitude to school (Enjoyment) by Year Level
The results summarised in Figure 5 show large gender differences in *Attentiveness* at all Grade/Year levels, with females being rated by their teachers as demonstrating significantly greater attentiveness than males. Note that all mean differences were statistically significant beyond the $p < 0.001$ level, by univariate two-tailed tests.
3. There are differences in ‘student ability’ and ‘intake characteristics’ at the class/teacher and school levels

It has already been noted that schools often vary significantly in the nature of their student intakes on account of socio-demographic factors and the existence of selective schools (p. 9). To illustrate these phenomena, data from the first two phases of the VQSP (1992 and 1993) are used here to explore variability within the sample in Student Ability and intake characteristics (Socio-educational Level) at the school, class/teacher and student levels. This was achieved by fitting multilevel variance components models to the relevant data for elementary and secondary students using MLn (Rasbash & Woodhouse, 1995, 1996). The general form of these models can be written as:

\[ y_{ijk} = a_{000} + (v_k + u_{jk} + e_{ijk}), \]

where:

- \( y_{ijk} \) = the response for student \( i \), in class \( j \), in school \( k \);
- \( a_{000} \) = an overall intercept term, constant across all students, classes and schools;
- \( v_k \) = a random coefficient for school \( k \);
- \( u_{jk} \) = a random coefficient for class \( j \) and school \( k \); and
- \( e_{ijk} \) = a random coefficient for student \( i \) in class \( j \) and school \( k \).

By fitting this model, it was possible to obtain estimates of the variances of the three random terms in equation [11], namely the variances of the school-, class- and student-level terms (i.e., \( \sigma^2_v \), \( \sigma^2_u \), and \( \sigma^2_e \), respectively), and to compute the proportion of residual variance due to differences between-schools \( \frac{\sigma^2_v}{\sigma^2_v + \sigma^2_u + \sigma^2_e} \), between-classes \( \frac{\sigma^2_u}{\sigma^2_v + \sigma^2_u + \sigma^2_e} \), and within students \( \frac{\sigma^2_e}{\sigma^2_v + \sigma^2_u + \sigma^2_e} \). Results of these analyses are summarised in Table 3.

Table 3 reveals important characteristics of the Victorian school system. In terms of student ability, there is little variation between schools at the elementary level, reflecting the relative homogeneous character of elementary schooling in Australia. At the secondary level, where 36 per cent of students are enrolled in fee-paying, non-Government, Catholic and independent schools (Victoria, 1993), there is significant variation among schools. For the present sample, 23.8 per cent of elementary students (Grades K to 6) and 31.5 per cent of secondary students (Years 7 to 12) were drawn from Catholic and independent schools. The between-classes variation, however, is quite small, reflecting the universal practice of randomly allocating students to mixed ability classes and the absence of ‘tracking’ or ‘streaming’.

In terms of students’ intake characteristics (Socio-educational Level), between-classes variation is non-existent, although there is substantial between-school variation, reflecting (in the case of elementary schools) the characteristics of local neighbourhoods on the socio-educational profiles of individual schools, and at the secondary level, both the influence of local neighbourhoods and the enrolment of significant proportions of students in fee-paying non-Government schools.

---

9 This is equivalent to equation [7] given on page 28. Note the change in notation for the intercept term (\( \gamma_{000} \)) to the simpler form, namely, \( a_{000} \).
Table 3. Summary Results of Fitting Three-Level Variance-Components Models: Residual Variance* in Student Ability and Socio-Educational Level - Within-Students, Among Classes and Between Schools

<table>
<thead>
<tr>
<th>Response Measure</th>
<th>Source of Variation</th>
<th>Within Students</th>
<th>Among Classes</th>
<th>Between Schools</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \sigma_{\text{se}}^2 ) (se)</td>
<td>( \sigma_{\text{uj}}^2 ) (se)</td>
<td>( \sigma_{\text{vk}}^2 )</td>
<td></td>
</tr>
<tr>
<td>Student ability</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td></td>
</tr>
<tr>
<td>(adjusted for Year level):</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elementary</td>
<td>2.39 (0.07)</td>
<td>0.23 (0.05)</td>
<td>0.18 (0.06)</td>
<td>6.4</td>
</tr>
<tr>
<td>Secondary</td>
<td>1.98 (0.07)</td>
<td>0.14 (0.04)</td>
<td>0.57 (0.20)</td>
<td>21.2</td>
</tr>
<tr>
<td>Socio-educational level:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elementary</td>
<td>0.80 (0.02)</td>
<td>0.00 (0.00)*</td>
<td>0.18 (0.04)</td>
<td>18.4</td>
</tr>
<tr>
<td>Secondary</td>
<td>0.78 (0.02)</td>
<td>0.01 (0.01)*</td>
<td>0.21 (0.07)</td>
<td>21.0</td>
</tr>
</tbody>
</table>

* Variance estimates with standard errors in parentheses. Unless otherwise indicated (*) all estimates are statistically significant at or beyond the \( p < 0.05 \) level by univariate two-tailed tests. Bolded percentages, from left to right, refer to the proportion of residual variance at the student-, class- and school-levels, respectively.

Not statistically significant.

4. Different adjustments result in different estimates of the proportion of variance in student achievement progress accounted for at the school, class/teacher and student levels

As indicated earlier, a major objective of the VQSP was to identify characteristics of schools that are effective, given their student intake characteristics, including initial levels of achievement (see p. 12). To achieve this, and given the inherent hierarchical structure of the data (i.e., students nested within class/teacher groups within schools), it was possible to estimate the relative importance of effects operating at the student-, class/teacher- and school-levels in accounting for variation in students' achievement progress in literacy and mathematics. It was also noted (see pp. 9-10) that is likely that different adjustments will result in different estimates of the proportion of variance in students' progress accounted for at each of these levels. If this is the case, it is important to identify clearly the particular kind of adjustment used, and where possible, to report results based on more than one kind of adjustment.

To investigate student progress using data obtained during phases 1 and 2 of the project, for example, three scores were calculated for each student. The first was the student’s achievement in literacy and mathematics at the end of 1993, as assessed by the teacher using the Victorian English and Mathematics Profiles, adjusted for the Grade/Year Level of the student and his or her Prior Achievement (assessed at the end of 1992). This constitutes a measure of residual change over a twelve month period. The second was the student’s achievement at the end of 1993 adjusted for Grade/Year Level, Prior
Achievement and any statistically significant intake characteristics (i.e., Gender, Socio-Educational Level, Non-English Speaking Background and whether the student lived in a Rural location). This is a measure of residual change further adjusted for student intake characteristics. The third was the student’s achievement at the end of 1993 adjusted for Grade/Year Level and Student Ability only. This measure is consistent with similar outcome measures used in many cross-sectional studies of school effectiveness in which achievement scores are adjusted for initial ability.

In calculating these three scores, a three-level regression model was fitted to the data and adjustments to the criterion variable (student achievement at the end of 1993) were made on the basis of the fixed parameter estimates from this model. Thus, for example, in the case of the first score calculated for Elementary literacy, the following model was fitted:

$$y_{ijk} = a_{000} + c_1 d_{ijk} + c_2 d_{2ijk} + b_{ijk} x_{ijk} + (v_k + u_{jk} + e_{ijk}),$$  \[12\]

in which:

- $y_{ijk}$ is the response variable (student achievement in literacy at the end of 1993) for student $i$, in class $j$, in school $k$;
- $a_{000}$ is an overall intercept term constant across all students, classes and schools;
- $c_1$ is the expected difference between students in Grades 1 and 3;
- $c_2$ is the expected difference between students in Grades 1 and 5;
- $d_{ijk}$ is a dummy variable coded 1 = student in Grade 3, 0 = otherwise;
- $d_{2ijk}$ is a dummy variable scored 1 = student in Grade 5, 0 = otherwise;
- $b_{ijk}$ is a slope coefficient for the regression of the response variable on the measure of Student Ability, random both between classes and between schools;
- $x_{ijk}$ is the Student Ability score for student $i$ in class $j$ in school $k$;
- $v_k$ is a random coefficient for school $k$;
- $u_{jk}$ is a random coefficient for class $j$ in school $k$; and
- $e_{ijk}$ is a random coefficient for student $i$ in class $j$, in school $k$.

The adjusted score ($y^*_{ijk}$) was then calculated as follows:

$$y^*_{ijk} = y_{ijk} - (c_1 d_{ijk}) - (c_2 d_{2ijk}) - (b_{ijk} x_{ijk}).$$  \[13\]

The same procedure was used to derive the additional, adjusted measures of student progress. The results of fitting the components models described by equations [12] and [13] to these three measures of student progress are shown in Table 4.

From Table 4 it is evident that all three measures are associated with high proportions of between-class/teacher variation, ranging from 38 to 64 per cent of the total variance. The third measure (student achievement adjusted for Grade/Year Level and Student Ability) is associated with the highest proportions of between-class/teacher variance. The between-school variance is relatively small, although somewhat higher in the case of the third measure of student progress in literacy for elementary students, namely the...
student’s literacy achievement at the end of 1993 adjusted for Grade/Year Level and Student Ability.

Table 4. Proportion of Residual Variance in Three Measures of Student Progress: Within-Students, Among Classes and Between Schools

<table>
<thead>
<tr>
<th>Student Progress</th>
<th>Within Students %</th>
<th>Among Classes %</th>
<th>Between Schools %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Elementary Literacy adjusted for:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Grade Level &amp; Prior Achievement</td>
<td>46.0</td>
<td>45.4</td>
<td>8.6</td>
</tr>
<tr>
<td>2. Grade Level, Prior Achievement &amp; intake</td>
<td>46.1</td>
<td>46.2</td>
<td>7.7</td>
</tr>
<tr>
<td>3. Grade Level, Student Ability</td>
<td>43.1</td>
<td>41.5</td>
<td>15.4</td>
</tr>
<tr>
<td><strong>Secondary Literacy adjusted for:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Grade Level &amp; Prior Achievement</td>
<td>54.9</td>
<td>37.8</td>
<td>7.3</td>
</tr>
<tr>
<td>2. Grade Level, Prior Achievement &amp; intake</td>
<td>56.4</td>
<td>38.1</td>
<td>5.5</td>
</tr>
<tr>
<td>3. Grade Level, Student Ability</td>
<td>49.6</td>
<td>43.7</td>
<td>6.7</td>
</tr>
<tr>
<td><strong>Elementary Mathematics adjusted for:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Grade Level &amp; Prior Achievement</td>
<td>41.8</td>
<td>53.6</td>
<td>4.6</td>
</tr>
<tr>
<td>2. Grade Level, Prior Achievement &amp; intake</td>
<td>41.2</td>
<td>55.4</td>
<td>3.4</td>
</tr>
<tr>
<td>3. Grade Level, Student Ability</td>
<td>37.3</td>
<td>53.6</td>
<td>9.1</td>
</tr>
<tr>
<td><strong>Secondary Mathematics adjusted for:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Grade Level &amp; Prior Achievement</td>
<td>38.8</td>
<td>52.8</td>
<td>8.4</td>
</tr>
<tr>
<td>2. Grade Level, Prior Achievement &amp; intake</td>
<td>39.4</td>
<td>53.5</td>
<td>7.1</td>
</tr>
<tr>
<td>3. Grade Level, Student Ability</td>
<td>36.0</td>
<td>64.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

5. Different adjustments also result in different estimates for explanatory variables having direct effects on students' achievement progress

To illustrate this finding, Table 5 summarises the results of fitting a three-level regression model to elementary students’ progress in Literacy, using two methods of adjustment, namely, (1) achievement at the end of 1993 adjusted for Grade/Year Level and Prior Achievement at the end of 1992, and (2) achievement at the end of 1993 adjusted for Grade/Year Level and Student Ability. The explanatory variables included in the model are listed in the first column. The notation $X_a$ has been used to denote explanatory variables measured at the student level, and $Z_a$ to denote explanatory variables measured at the class/teacher level. No variables measured at the school level were found to be significant in explaining student progress. To assist interpretation, the explanatory variables listed in column one have each been re-scaled to zero mean and unit standard deviation. Hence, the parameter estimates ($b$ est) in columns two and three are standardized coefficients, each of which (in unstandardized metric) is at least twice its standard error.
Table 5. Explanatory Models of Student Progress Fitted to Two Adjusted Measures of Elementary Student Progress in Literacy: Variance Estimates ($\hat{b}$ est)\(^a\) with Standard Errors (in parentheses)

<table>
<thead>
<tr>
<th>Explanatory Variables</th>
<th>Elementary Literacy adjusted for Grade level and Prior Achievement ($\hat{b}$ est (se))</th>
<th>Elementary Literacy adjusted for Grade level and Student Ability ($\hat{b}$ est (se))</th>
</tr>
</thead>
<tbody>
<tr>
<td>($X_n$ Student-Level Variables)</td>
<td>($Z_m$ Class/Teacher-Level Variables)</td>
<td>($X_n$ Student-Level Variables)</td>
</tr>
<tr>
<td>$a_{000}$ Constant</td>
<td>-0.198 (0.11)</td>
<td>0.124 (0.10)</td>
</tr>
<tr>
<td>$X_1$ Socio-educational level</td>
<td>0.062 (0.01)</td>
<td>0.040 (0.01)</td>
</tr>
<tr>
<td>$X_2$ Gender (Female)</td>
<td>---</td>
<td>0.127 (0.03)</td>
</tr>
<tr>
<td>$X_3$ Non-English Speaking Background</td>
<td>---</td>
<td>-0.228 (0.05)</td>
</tr>
<tr>
<td>$X_4$ Attentiveness(^b)</td>
<td>0.271 (0.01)</td>
<td>0.194 (0.02)</td>
</tr>
<tr>
<td>$Z_1$ Literacy In-Service Programs</td>
<td>0.406 (0.11)</td>
<td>---</td>
</tr>
<tr>
<td>$Z_2$ Composite Class</td>
<td>-0.271 (0.09)</td>
<td>-0.330 (0.12)</td>
</tr>
<tr>
<td>$Z_3$ Work Demands</td>
<td>-0.096 (0.04)</td>
<td>-0.103 (0.05)</td>
</tr>
</tbody>
</table>

**Per Cent of Variance Explained**

| 23.7 | 24.6 |

**Random:**

- **Between-schools variance**
  - Variance of intercepts - $\sigma^2_{(30)}$ 0.046 (0.025)* 0.141 (0.052)

- **Between-classes variance**
  - Variance of intercepts - $\sigma^2_{(20)}$ 0.408 (0.041) 0.380 (0.050)
  - Variance of slopes - $\sigma^2_{(21)}$ 0.021 (0.004) 0.040 (0.008)
  - Intercept-slope covariance - $\sigma_{(201)}$ 0.040 (0.010) 0.047 (0.015)

- **Corr. of slopes and intercepts - $\rho$** 0.432 0.381

- **Within Students variance**
  - $\sigma^2_{(10)}$ or $\sigma^2_{eijk}$ 0.345 (0.009) 0.341 (0.011)

---

\(^a\) Unless otherwise indicated (*), all $\hat{b}$ coefficients are statistically significant at or beyond the $p < 0.05$ $\alpha$ level by univariate two-tailed tests.

\(^b\) Random at the class/teacher level

From Table 5, the model explains similar proportions of the variance for the two methods of adjustment for student progress in **Literacy**. Nevertheless, the magnitudes of the coefficients indicate that the effects of particular explanatory variables differ depending on the method of adjustment. In particular, **Gender** and **Non-English Speaking Background** have no significant effect when measures of student achievement...
are adjusted for Grade Level and Prior Achievement, but Gender has a significant positive effect and Non-English Speaking Background has a significant negative effect on measures of student achievement when adjusted for Grade Level and Student Ability. In addition, the extent to which teachers participate in Literacy In-Service Programs has a significant effect on the first but not on the second of the two adjusted measures of student progress. However, for both adjustment methods the ratio of the variance estimates for the class/teacher slopes (\( \sigma^2_{(a)} \) - for the random effect of Attentiveness) to their respective standard errors are significant, indicating that the relationship between student progress and Attentiveness varies significantly between classes. Further, the magnitudes of the correlations between the slopes and intercepts indicate a moderate tendency for larger boosts in student progress in those classes having higher mean levels of Attentiveness in the classroom.

6. The key to improved educational outcomes is teacher effectiveness

Analyses of the data summarised in Tables 4 and 5 cast new light on the findings of much previous research and emphasise the importance in studies of school effectiveness which enable class/teacher effects to be investigated (see Hill & Rowe 1996; Rowe et al., 1995; Scheerens, 1993). In commenting on the findings reported by Scheerens, Vermeulen and Pelgrum (1989), Scheerens (1993) observes: "...teacher and classroom variables account for more of the variance in pupil achievement than school variables. Also, in general, more powerful classroom level variables are found that account for between-class variance than school level variables in accounting for between-school variance" (p. 20). In the light of this observation, it is somewhat curious that the twin disciplines of school effectiveness and school improvement (see Reynolds, Hopkins & Stoll, 1993) have focussed and continue to focus predominantly on school level organisational factors rather than on factors at the class/teacher level.

Above all, these findings underscore the fact that learning takes place in classrooms through the interaction of students and their teachers. Interpreting these findings requires some care, however, since they have a number of possible explanations. One is that they reflect streaming of classes. From the results presented in Table 3, this explanation can almost certainly be discounted and underscores the fact that mixed ability classes have for some time been the universal norm in Victorian schools. A second explanation is that these results reflect inconsistent use of the Profiles by teachers. Here again, there is little evidence in support of this explanation, as the Profiles have been demonstrated to have high internal consistency and inter-rater reliability (see Tables 1 and 2). In addition, efforts were made to identify and remove from the analyses atypical and unexpected teacher ratings of which there were very few.

A third explanation for the finding of large class/teacher effects and small to insignificant school effects, is that it primarily reflects variations in teacher quality. The suggestion here is that it is essentially through the quality of teaching that effective schools make a difference. In fact, on the basis of findings from the project it could be argued that effective schools are only 'effective' to the extent that they have effective teachers. This outcome is consistent with the findings of a study of Victorian primary schools by Ainley, Goldman & Reid (1990) which found that differences among teachers within schools were greater than differences between schools in the growth they achieved in their students. Moreover, from a recent Australian national survey of community views...
of "What makes an effective school?", McGaw, Piper, Banks and Evans (1992) found that the most frequently mentioned factor was "the quality of the teachers" (Piper, 1993, p. 2), constituting 65 per cent of all responses.

7. The effects of particular variables in explaining students' progress in achievement vary across subjects and across elementary and secondary schools

Table 6 summarises results of fitting three-level regression models to four sets of data, namely to a measure of student progress in Literacy and Mathematics for both elementary and secondary students.

Table 6. Explanatory Models for Elementary and Secondary Students' Progress in Literacy and Mathematics Using Measures Adjusted for Year Level and Prior Achievement, Showing Standardized Coefficients (b)\(^a\)

<table>
<thead>
<tr>
<th>Explanatory Variables</th>
<th>Literacy</th>
<th>Mathematics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Elementary</td>
<td>Secondary</td>
</tr>
</tbody>
</table>

\begin{tabular}{lccccc}
\hline
Fixed: & & & & \\
X_1 Socio-educational level & 0.062 & 0.034 & 0.039 & — \\
X_2 Attentiveness & 0.271 & 0.379 & 0.266 & 0.345 \\
X_3 Gender (Female) & — & 0.090 & -0.144 & -0.179 \\
X_4 Attitude to School & — & 0.034 & — & 0.045 \\
X_5 Homework & — & 0.097 & — & — \\
X_6 Perceived Usefulness & — & — & — & 0.042 \\
Z_1 P.D. In-Service Programs & 0.406 & 0.258 & 0.416 & — \\
Z_2 Composite Class & -0.271 & — & — & — \\
Z_3 Work Demands & -0.096 & — & — & — \\

Random: & & & & \\
Between-schools variance & & & & \\
Variance of intercepts - \(\sigma^2_{(j0)}\) & 0.046* & 0.059* & 0.035* & 0.077* \\

Between-classes variance & & & & \\
Variance of intercepts - \(\sigma^2_{(2j)}\) & 0.408 & 0.354 & 0.486 & 0.538 \\
Variance of slopes - \(\sigma^2_{(2i)}\) & 0.021 & 0.034 & 0.023 & 0.040 \\
Interc.-slope covariance - \(\sigma_{(2j0)}\) & 0.040 & 0.016* & 0.024 & 0.019* \\
Corr. of sl. & interc. - \(p\) & 0.432 & 0.146 & 0.227 & 0.130 \\

Within Students variance & & & & \\
\(\sigma^2_{(1j)}\) or \(\sigma^2_{eijk}\) & 0.345 & 0.382 & 0.311 & 0.288 \\
\hline
\end{tabular}

\(a\) Unless otherwise indicated (*) all \(b\) coefficients are statistically significant at or beyond the \(p < 0.05\) level by univariate two-tailed tests.

\(b\) Random at the class/teacher level.

Centre for Applied Educational Research, The University of Melbourne
In each case, the measure of Student Progress was student achievement adjusted for Grade Level and Prior Achievement. It is apparent that the effects of particular explanatory variables are not always the same across subjects and across elementary and secondary schools. For example, when a student is placed in a Composite Class made up of students from two or more Grade/Year levels, the effect is significantly negative in the case of elementary Literacy, but not in the case of elementary Mathematics. (Note that there were no composite classes among secondary schools in the VQSP). The effect of Gender on student progress is negative for Mathematics at both the elementary and secondary levels, with girls making less progress than their male counterparts, but is positive for Literacy. The effect of teacher participation in specialist Literacy In-Service Programs is substantial for elementary students and to a lesser extent for secondary students with respect to Literacy, but is not significant for secondary Mathematics.

8. Attentiveness has a massive effect on student achievement progress

The functional relationship between academic under-achievement and behaviour problems of an externalizing or 'acting-out' kind is well documented, especially with measures of reading achievement. What is not so well documented are reliable estimates of the proportion of variation of students' achievements in numeracy (mathematics) and in literacy (reading, writing and spoken language) due to such behaviours. Despite large and significant gender differences in students' externalizing behaviours in the classroom such as Irritatable/Antisocial, Inattentive and Restless (eg., Figure 5, p. 39), teachers have generally positive perceptions of the behaviour of students, with negative behaviour being seen as characteristic of a minority of students.

In the context of fitting explanatory models, however, it is Attentiveness/Inattentiveness in the classroom that is the most consistent and salient factor in explaining both elementary and secondary students' achievement progress (see Table 6), ranging from standardised effect coefficients of 0.266 (for elementary students' progress in Mathematics) to 0.379 (for secondary students' progress in Literacy). This finding is commensurate with a large clinical research literature showing strong relationships between students' learning outcomes and especially their attentive/inattentive behaviours (see references cited in Note 10). For illustrative purposes, Figures 6.1 and 6.2 summarise key findings from the first phase of the VQSP, as documented by Rowe and Rowe (1993).

The data presented graphically in Figures 6.1 and 6.2 derive from a series of analyses in which students' total Literacy scores (for reading, writing and spoken language) and Mathematics scores (for number and space) – standardised within each age sub-group and adjusted for SEL – were regressed on the respective item scores within each scale of the 16-item Teacher Form of the Rowe Behavioural Rating Inventories (RBRI, see p. 19). These analyses used the polyserial correlation estimates; that is, the continuous values for Literacy and Numeracy, and ordinal values for the constituent item scores (see Olsson, Drasgow & Dorans, 1982).

Figure 6.1 Percentage histogram showing proportions of explained variation (unique) in achievement due to RBRI scales, for males and females in elementary schools.

Figure 6.2 Percentage histogram showing proportions of explained variation (unique) in achievement due to RBRI scales, for males and females in secondary schools.
Above all, the findings shown in Figures 6.1 and 6.2 underscore the importance of taking into account students’ externalizing behaviours in the classroom (RBRI score) and especially Attentiveness, in the provision of any explanation of factors affecting student learning outcomes. Moreover, in the study reported by Rowe and Rowe (1992b), non-recursive structural equation modelling revealed significant interdependent effects (reciprocal) between students’ reading achievements and inattentiveness in the classroom, suggesting the need for intervention strategies to focus on both domains simultaneously.

Whereas the higher attentiveness levels of girls is familiar to most teachers (see Figure 5), the implications for school-based learning and its assessment may not always be recognised. In recent years, there has been a greater emphasis within Australian schools, both in approaches to teaching and learning and to assessment of student achievement, on activities that require high levels of sustained attention. Such activities include on-task-demanding behaviours such as the production of written portfolios, the writing of extended pieces of prose, and the completion of written research projects such as common assessment tasks (CATs). There has been a corresponding move away from short answer and ‘check the box’ type assessment activities to tasks requiring increasingly higher levels of verbal reasoning skills – activities in which girls have a well-established achievement and maturational advantage. It is possible that these changes in pedagogy may have placed, albeit inadvertently, a greater premium on attentiveness that have contributed to the phenomenon of substantial gender differences in students’ achievement progress, mediated especially through Attentiveness (see Hill & Rowe, in press; Rowe, 1991b, 1995; Rowe & Rowe, 1992b, and Figure 7 below).

While it is not possible to comment in detail here about the underlying causes of inattentiveness, or the extent to which attentiveness can be enhanced by the teacher and the school, there are grounds for believing that in the majority of cases psychosocial factors such as anxiety or stress may be the most common explanation for students’ inattentive behaviours. For a minority of students, inattentiveness may be related to auditory or other perceptual problems, or to deeper-seated specific learning difficulties. Nonetheless, these are important issues requiring further investigation.

9. There are complex, interdependent effects among explanatory variables that account for students’ achievement progress

The results of the fitted models summarised in Table 6 are limited to variables that have direct effects on student progress. Most models of school effectiveness, however, typically assume more complex structural relationships between networks of variables operating at different levels. The approach taken here to exploring such relationships was to form a set of conditional structural equations that could be estimated in the same way as ordinary path models, that is, fitting each equation separately, but using three-level regression analysis to take into account the hierarchical structure of the data with students nested within classes/teachers within schools. As discussed by Bosker and Scheerens (1994, p. 169), such an approach is not ideal, but is useful as a means of undertaking exploratory analyses of indirect, ‘putative’ causal effects. When undertaking confirmatory analyses it is desirable that the set of equations be estimated jointly. Where the response variables in a set of equations are all at the same level, this can be accomplished in the software package MLn (Rasbash & Woodhouse, 1995, 1996)
through the specification of the set of equations as a multivariate, multilevel regression model (Hill & Rowe, in preparation).

Table 7 presents a set of four structural equations for elementary student progress in Literacy, while Figure 7 depicts these equations graphically as a multi-level path analysis diagram for student Progress in Literacy.

Table 7. Set of Four Multi-Level Structural Equations for Elementary Students' Progress in Literacy\(^\text{a}\)\(^\text{b}\) Showing Standardized Path Coefficients ($b$)\(^\text{b}\)

<table>
<thead>
<tr>
<th>Equation</th>
<th>Explanatory Variables</th>
<th>$b$</th>
<th>se</th>
<th>Equation</th>
<th>Explanatory Variables</th>
<th>$b$</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eq. 1: Progress in Literacy</td>
<td>$X_1$ Socio-educational level</td>
<td>.062</td>
<td>.012</td>
<td>$X_2$ Attentiveness</td>
<td>.271</td>
<td>.014</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$Z_1$ Literacy/PD Programs</td>
<td>.406</td>
<td>.105</td>
<td>$Z_2$ Composite Classes</td>
<td>-.271</td>
<td>.087</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$Z_3$ Work Demands</td>
<td>-.096</td>
<td>.041</td>
<td>$X_3$ Gender (Female)</td>
<td>.125</td>
<td>.023</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$X_4$ Non-Eng. Speaking NESB</td>
<td>.600</td>
<td>.017</td>
<td>$Z_4$ Teacher Affect</td>
<td>.038</td>
<td>.017</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$X_5$ Homework</td>
<td>.143</td>
<td>.013</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$X_6$ Teacher Responsiveness</td>
<td>.248</td>
<td>.044</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$X_7$ Rural</td>
<td>.092</td>
<td>.022</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$X_8$ Teacher Responsiveness</td>
<td>.130</td>
<td>.050</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$Z_4$ Teacher Affect</td>
<td>.077</td>
<td>.033</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For data such as those described here, there are two key advantages in fitting unconditional, confirmatory, structural equation models (SEM) of the EQS kind (Bentler & Wu, 1993) or LISREL kind (Jöreskog & Sörbom, 1993a). First, they provide a means of estimating jointly the magnitude of direct, indirect and interdependent effects among variables. Second, they can account for measurement error in both the observed and latent variables. However, there are major limitations and disadvantages. Such SEM approaches assume that the sample variables, regardless of level, are independently distributed in a multivariate population. That is, an unconditional SEM model assumes single-level data. Hence, short of applying procedures for 'purging' computed variance-covariance matrices of the assumptions of independence, there is no simple adjustment of the structural modeling framework that can be made in order to deal with non-independent observations in multilevel data of the present kind (see Cuttance, 1987, pp. 250, 263; Scheerens, 1993, p. 31).

While it is possible to model a few groups separately (e.g., classes/teachers, or schools) in a multi-group structural equation model, such an approach becomes intractable with more than a small number of groups (i.e., a maximum of 10 in LISREL 8). The crucial drawback of fitting such models is that they do not allow for reliable modeling of substantive relationships across levels. Moreover, such models typically lack power due to the small number of observations in the higher-level units. In the context of school effectiveness research, Keeskes and Cheung (1990) note that attempts to apply structural equation models of the EQS and LISREL type "...can be said to be an inelegant or clumsy approach that does not lead to an effective analysis of the hierarchically structured data" (p. 275).

Thus, given both the sampling structure of the present data and the large proportions of computed residual variance in the response variables at the class/teacher level, it was decided that the advantages favoured the use of multi-level path analysis.

Centre for Applied Educational Research, The University of Melbourne
Footnote to Table 7:

a Progress in Literacy was measured as Literacy achievement adjusted for Grade Level and Prior Achievement and with the variance of Prior Achievement specified as random at the class/teacher and school levels. Moreover, as a predictor of Progress in Literacy, the variance of Attentiveness was specified as random at the class/teacher level.

b All path coefficients (b) are statistically significant at or beyond the p < 0.05 α level by univariate two-tailed tests.

---

Figure 7. Multi-level path model of elementary student progress in Literacy showing standardised path coefficients
(Note: all coefficients are significant at the p < 0.05 level, by univariate two-tailed tests)

The results summarised in Table 7 and Figure 7 indicate that the main effects on Progress in Literacy (with adjustments for Grade Level and Prior Achievement) for elementary students are: student Attentiveness in the classroom, whether the teacher has participated in specialist Literacy In-Service Programs, and whether the student is placed in a Composite Class containing students from two or more Grade levels. Attentiveness and Literacy In-Service Programs are positive in their effect whereas being in a Composite Class has a strong negative effect.

Attentiveness is in turn explained by a number of variables including Socio-Educational Level and by those students residing in Rural locations, but especially by the Gender of the student, with female students being more attentive than males. Somewhat surprisingly, Attitude to School (i.e., Enjoyment) is weakly related to Attentiveness and has no direct effect on elementary students' Progress in Literacy. It is, however,
positively influenced by *English Homework* and strongly influenced by students’ perceptions of *Teacher Responsiveness* and whether they are from a *Non-English Speaking Background (NESB)*, with NESB students having a more positive *Attitude to School* than their English-speaking counterparts.

10. **Teachers’ affect is high and perceptions of their work environments are generally positive, but there are significant differences between the perceptions of elementary and secondary teachers, as well as large between-school differences**

Using a semantic differential instrument consisting of items requiring responses on seven-point evaluative scales (see p. 21), two *Teacher Affect* scales were used in the VQSP, namely, *Energy/Enthusiasm* and *Warmth towards students*. Composite scores for these scales were computed as described on pages 22-26. To assist interpretation, the composite scores were recoded from 1 to 7, to -3 to +3. For illustrative purposes, Figure 8.1 summarises the *Affect* data on two scales for elementary (primary) and secondary teachers ($n_p = 574$ and $n_s = 723$, respectively) from the 1994 data collection phase in the form of ‘box-and-whisker’ plots.

![Figure 8.1 Box plots showing the distribution of primary and secondary teachers’ responses on two Affect scales](image)

Figure 8.1 indicates that all teachers’ *Affect* levels are high, but that elementary (primary) teachers report notably higher levels of *Energy/Enthusiasm* and *Warmth towards Students* than their secondary counterparts. Similar patterns were also observed for teachers’ responses on the 12 scales related to perceptions of their work environment (recoded from 1 to 5, to -2 to +2), as shown in Figure 8.2. (For specific details of the source of these scales, see p. 21).
Figure 8.2 Box plots showing the distribution of elementary (primary) and secondary teachers' responses on 12 Work Environment scales.
From Figure 8.2, apart from perceived Work Demands, the distributions for elementary (primary) teachers' responses on each of the work environment scales are notably (and significantly) more positive than those for secondary teachers. Nonetheless, the consistent positive ratings by all teachers is of interest at a time when it is often asserted that teacher 'morale is low'. The exception was teacher ratings of the amount of Feedback they receive on their work performance (especially by secondary teachers), suggesting that systematic consideration of issues associated with teacher appraisal is long overdue.

The findings presented in Figures 8.1 and 8.2, however, provide only a limited account of teachers' Affect and perceptions of their work environment. There were also large and significant differences in the proportion of variance in these scales due to the multilevel structure of the data; that is, teachers grouped within schools – particularly for the work environment scales. To illustrate these differences, Table 8 summarises the results of fitting simple, two-level, variance components models to each of the affect and work environment scale scores for the 1994 primary and secondary teacher sample.

<table>
<thead>
<tr>
<th>Affect and Work Environment Scales</th>
<th>Primary Teachers</th>
<th>Secondary Teachers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Within Teachers %</td>
<td>Between Schools %</td>
</tr>
<tr>
<td>Energy/Enthusiasm</td>
<td>96.7</td>
<td>3.3</td>
</tr>
<tr>
<td>Warmth towards students</td>
<td>97.5</td>
<td>2.5</td>
</tr>
<tr>
<td>Morale</td>
<td>75.7</td>
<td>24.3</td>
</tr>
<tr>
<td>Feedback</td>
<td>88.0</td>
<td>12.0</td>
</tr>
<tr>
<td>Work Demands</td>
<td>79.6</td>
<td>20.4</td>
</tr>
<tr>
<td>Leadership Support</td>
<td>83.3</td>
<td>16.7</td>
</tr>
<tr>
<td>Discipline Policy</td>
<td>69.3</td>
<td>31.7</td>
</tr>
<tr>
<td>Peer Support</td>
<td>80.6</td>
<td>19.4</td>
</tr>
<tr>
<td>Role Clarity</td>
<td>87.6</td>
<td>12.4</td>
</tr>
<tr>
<td>Professional Development</td>
<td>82.7</td>
<td>17.3</td>
</tr>
<tr>
<td>Goal Congruence</td>
<td>84.2</td>
<td>15.8</td>
</tr>
<tr>
<td>Decision Making</td>
<td>86.5</td>
<td>13.5</td>
</tr>
<tr>
<td>Student Orientation</td>
<td>85.2</td>
<td>14.8</td>
</tr>
<tr>
<td>Curriculum Co-ordination</td>
<td>82.7</td>
<td>17.3</td>
</tr>
</tbody>
</table>

It should be noted that while these data derive from the third data collection phase (1994) of the VQSP, the distributions on these scales obtained from the 1992 and 1993 phases were almost identical to those from 1994 as reported here.
11. Leadership support is critically important in establishing a positive work environment for teachers

To examine the magnitude of the interdependent effects among the Affect and work environment scale scores, a single-level, multiple-group, structural equation model was fitted simultaneously to the relevant data for the 1992, 1993 and 1994 teacher groups. The model specifications for this analysis have been given in equations [8] to [10] on page 30. Figure 9 summarises the results of this analysis and Table 9 presents the direct, indirect and total effects.

Figure 9. Multi-Group model showing direct effects among teachers' perceptions of themselves and their work environments: Completely standardized, common metric path coefficients estimated simultaneously from the 1992, 1993 and 1994 samples.* All paths are statistically significant at the p < .001 level by univariate two-tailed tests. (Percentages refer to proportion of variance due to between-school differences)

*1992: 376 primary teachers in 59 schools; 555 secondary teachers in 30 schools
*1993: 480 primary teachers in 51 schools; 630 secondary teachers in 28 schools
*1994: 574 primary teachers in 50 schools; 723 secondary teachers in 25 schools

In analysing data from several samples simultaneously, according to models for each group with some or all of the parameters constrained to be equal over groups, it is essential that covariance matrices are used since one cannot make legitimate comparisons of the estimated parameters from correlation matrices. Moreover, to compare parameters across groups, the variables must be measured in a common metric for all groups. For these reasons, "It is important to understand that in a multi-sample analysis with constraints across groups, one must not standardize the variables within each group. Neither the observed nor the latent variables should be standardized" (Jöreskog & Sörbom, 1989b, p. 238).
From Figure 9, the 'goodness of fit' indices indicate an excellent fit of the model to the data. The magnitudes of the standardized path coefficients are interpreted as (for example) “one standard deviation increase in Leadership Support ‘leads to’ 0.899 of a standard deviation increase in Goal Congruence”. In fact, the entire model is dominated by the direct, indirect and total effects of Goal Congruence, encouragement to undertake Professional Development, and especially Leadership Support.

To further illustrate this, Table 9 gives the standardized direct, indirect and total effects of Leadership Support, Goal Congruence and Professional Development on other affect and work environment dimensions.

Table 9. Standardized Direct, Indirect and Total Effects

<table>
<thead>
<tr>
<th>Leadership Support on:</th>
<th>Direct Effect</th>
<th>Indirect Effect</th>
<th>Total Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy/Enthusiasm</td>
<td>.374</td>
<td>.374</td>
<td></td>
</tr>
<tr>
<td>Warmth towards students</td>
<td>.260</td>
<td>.260</td>
<td></td>
</tr>
<tr>
<td>Morale</td>
<td>.836</td>
<td>.836</td>
<td></td>
</tr>
<tr>
<td>Feedback</td>
<td>.539</td>
<td>.243</td>
<td>.782</td>
</tr>
<tr>
<td>Work Demands</td>
<td>-.216</td>
<td>-.216</td>
<td></td>
</tr>
<tr>
<td>Discipline Policy</td>
<td>.715</td>
<td>.715</td>
<td></td>
</tr>
<tr>
<td>Peer Support</td>
<td>.827</td>
<td>.827</td>
<td></td>
</tr>
<tr>
<td>Role Clarity</td>
<td>.713</td>
<td>.143</td>
<td>.856</td>
</tr>
<tr>
<td>Professional Development</td>
<td>.832</td>
<td></td>
<td>.832</td>
</tr>
<tr>
<td>Goal Congruence</td>
<td>.899</td>
<td></td>
<td>.899</td>
</tr>
<tr>
<td>Decision Making</td>
<td>.880</td>
<td></td>
<td>.880</td>
</tr>
<tr>
<td>Student Orientation</td>
<td>.763</td>
<td>.763</td>
<td></td>
</tr>
<tr>
<td>Curriculum Coordination</td>
<td>.833</td>
<td></td>
<td>.833</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Goal Congruence on:</th>
<th>Direct Effect</th>
<th>Indirect Effect</th>
<th>Total Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morale</td>
<td>.612</td>
<td></td>
<td>.612</td>
</tr>
<tr>
<td>Discipline Policy</td>
<td>.795</td>
<td></td>
<td>.795</td>
</tr>
<tr>
<td>Role Clarity</td>
<td>.159</td>
<td></td>
<td>.159</td>
</tr>
<tr>
<td>Student Orientation</td>
<td>.701</td>
<td></td>
<td>.701</td>
</tr>
<tr>
<td>Curriculum Coordination</td>
<td>.617</td>
<td></td>
<td>.617</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Professional Development on:</th>
<th>Direct Effect</th>
<th>Indirect Effect</th>
<th>Total Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy/Enthusiasm</td>
<td>.392</td>
<td>.058</td>
<td>.450</td>
</tr>
<tr>
<td>Warmth towards students</td>
<td>.226</td>
<td></td>
<td>.226</td>
</tr>
<tr>
<td>Morale</td>
<td>.030</td>
<td></td>
<td>.030</td>
</tr>
<tr>
<td>Feedback</td>
<td>.292</td>
<td></td>
<td>.292</td>
</tr>
<tr>
<td>Work Demands</td>
<td>-.260</td>
<td></td>
<td>-.260</td>
</tr>
<tr>
<td>Student Orientation</td>
<td>.159</td>
<td></td>
<td>.159</td>
</tr>
</tbody>
</table>

Note all effect coefficients are statistically significant at or beyond the $p < 0.05$ level, by univariate two-tailed tests.

Centre for Applied Educational Research, The University of Melbourne
The results presented in Figure 9 and Table 9 point to the following conclusions:

- **Leadership Support** is overwhelmingly important in establishing a positive working environment for teachers. In particular, it is associated with powerful direct effects on Feedback, teacher involvement in Professional Development activities, on teachers' perceptions of the amount of Peer Support they receive, on the degree of Goal Congruence among staff within the school, on their involvement in Decision Making and on their Role Clarity perceptions. In fact, Leadership Support has significant direct and/or indirect effects on all measured aspects of teachers' Affect and perceptions of their work environment.

- Teachers' perceptions of the degree of Goal Congruence within the school have strong direct effects on Morale, Student Orientation and Curriculum Co-ordination within the school, as well as having a considerable effect on teachers' perceptions of the school's Discipline Policy.

- Participation in Professional Development activities has a positive effect on teachers' levels of Energy/Enthusiasm, which in turn affects their feelings of Warmth towards students. It is also associated with positive effects on their perceptions of the amount of Feedback they receive and Student Orientation. Of special interest is the finding that participation in Professional Development activities leads to a significant decline in teachers' perceptions of Work Demands placed upon them.

12. **Teacher Morale is principally influenced by the direct and indirect effects of Leadership Support, Peer Support, Professional Development and Goal Congruence operating at both the school and teacher levels**

Of the various indicators of the quality of teachers' work environment, considerable interest surrounds that of teacher morale. To estimate the most important direct and indirect effects on teacher Morale, a series of multilevel, structural equation models were fitted to the total teacher data. The model providing the best explanation for Morale is summarised in Figure 10, with specific details of effects given in Table 10.

![Figure 10. Completely standardized direct effects for multilevel structural equation model for teachers' Morale, showing school-level estimates (above) and teacher-level estimates (below) * Not significant at the $p < 0.05$ level](image)

*Model Goodness-of-fit indices: $\chi^2 = 3.35$, $p = 0.187$; RMSEA = 0.015; GFI = 0.991; RMR = 0.002; CFI = 1.00; RFI = 0.996*

Centre for Applied Educational Research, The University of Melbourne
Table 10. Maximum Likelihood Solution to Multilevel Structural Equation Model\textsuperscript{14} Showing Standardized Direct, Indirect and Total Effects at the School-Level (above) and Teacher Level (below)

<table>
<thead>
<tr>
<th>Leadership Support on</th>
<th>Direct Effect</th>
<th>Indirect Effect</th>
<th>Total Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morale</td>
<td>.125</td>
<td>.569</td>
<td>.694</td>
</tr>
<tr>
<td></td>
<td>.161</td>
<td>.569</td>
<td>.730</td>
</tr>
<tr>
<td>Peer Support</td>
<td>.009*</td>
<td>.553</td>
<td>.562</td>
</tr>
<tr>
<td></td>
<td>.227</td>
<td>.429</td>
<td>.656</td>
</tr>
<tr>
<td>Goal Congruence</td>
<td>.457</td>
<td>.221</td>
<td>.678</td>
</tr>
<tr>
<td></td>
<td>.591</td>
<td>.130</td>
<td>.721</td>
</tr>
<tr>
<td>Professional Development</td>
<td>.486</td>
<td>–</td>
<td>.486</td>
</tr>
<tr>
<td></td>
<td>.643</td>
<td>–</td>
<td>.643</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Peer Support on:</th>
<th>Direct Effect</th>
<th>Indirect Effect</th>
<th>Total Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morale</td>
<td>.155</td>
<td>–</td>
<td>.155</td>
</tr>
<tr>
<td></td>
<td>.288</td>
<td>–</td>
<td>.288</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Goal Congruence on:</th>
<th>Direct Effect</th>
<th>Indirect Effect</th>
<th>Total Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morale</td>
<td>.710</td>
<td>.095</td>
<td>.806</td>
</tr>
<tr>
<td></td>
<td>.526</td>
<td>.110</td>
<td>.636</td>
</tr>
<tr>
<td>Peer Support</td>
<td>.616</td>
<td>–</td>
<td>.616</td>
</tr>
<tr>
<td></td>
<td>.382</td>
<td>–</td>
<td>.384</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Professional Development on:</th>
<th>Direct Effect</th>
<th>Indirect Effect</th>
<th>Total Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morale</td>
<td>–</td>
<td>.410</td>
<td>.410</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>.197</td>
<td>.197</td>
</tr>
<tr>
<td>Peer Support</td>
<td>.279</td>
<td>.280</td>
<td>.559</td>
</tr>
<tr>
<td></td>
<td>.239</td>
<td>.077</td>
<td>.316</td>
</tr>
<tr>
<td>Goal Congruence</td>
<td>.455</td>
<td>–</td>
<td>.455</td>
</tr>
<tr>
<td></td>
<td>.202</td>
<td>–</td>
<td>.202</td>
</tr>
</tbody>
</table>

\* Not statistically significant at the $p < 0.05$ level by univariate two-tailed test

\textsuperscript{14} This solution was based on a two-step procedure. First, a multivariate multilevel model was fitted to the five composite work environment variables to estimate the variance-covariance matrices, adjusted for Teacher Type (elementary and secondary) and Teacher Gender (male and female), at the school and teacher levels. The purpose of this approach was to partition the variances and covariances among the five composites into separate school- and teacher-level variance-covariance matrices, as a prelude for the second step which involved fitting an explanatory two-level (teacher and school) structural equation model. Specific details of these procedures for the present data are given by Rowe (1996) and by Rowe and Hill (in press).

Centre for Applied Educational Research, The University of Melbourne
The results shown in Figure 10 and Table 10 indicate significant direct effects on teachers' Morale (MORALE) from Leadership Support (LEADSUP), Peer Support (PEERSUP) and Goal Congruence (GOALCON), and indirect effects from Professional Development (PROFDEV), mediated by PROFDEV, PEERSUP and GOALCON. Moreover, there are indirect effects from GOALCON on MORALE, mediated by PEERSUP. However, a substantive interpretation of the solution suggests at least two notable features of the model fitted simultaneously at the school and teacher levels.

First, the direct effects of Goal Congruence (GOALCON) on Morale (MORALE) and Peer Support (PEERSUP), and Professional Development (PROFDEV) on Peer Support (PEERSUP) and Goal Congruence (GOALCON) are notably stronger at the school level than at the teacher level. In fact, at the school level, the magnitudes of the total effect estimates of Professional Development (PROFDEV) on Morale and Goal Congruence are more than twice those at the teacher level. Second, the salient feature of the solution is that valuable information at both the teacher and school levels can be obtained by accounting for the hierarchical structure of the data. Furthermore, the evidence from the model 'goodness-of-fit' indices (particularly, the RMSEA) suggest a high likelihood in the stability of these interdependent effects in accounting for variation in teachers' Morale across additional samples of teachers drawn from similar populations.

Above all, the finding regarding the overwhelming importance of Leadership support in establishing a positive work environment for teachers is of special significance given local and international trends toward greater self-governing autonomy for schools (Caldwell & Spinks, 1988; 1992). In so far as it is possible to predict 71 per cent of the variation in teacher Morale in terms of leadership support, goal congruence and peer support, it is clear that improving the quality of leadership support in the school is central to the establishment of positive teacher work environments (see Hill & Russell, 1993; Lee, Dedrick & Smith, 1991).

Taken together with the finding regarding the importance of Professional Development in increasing teachers' Energy, Warmth towards students and perceptions of Feedback, and on decreasing their perceptions of Work Demands, the teacher response data have yielded fruitful indications of how schools and school systems might usefully direct their efforts to improve teacher quality. Systematic professional development programs for teachers and leadership training programs for principals, vice-principals and senior staff stand out as areas worthy of close attention (see Ashton & Webb, 1986; Joyce & Showers, 1988; Rowe, 1990, 1995; Rowe & Sykes, 1989).

13. Modelling student progress using the three waves of data indicated that none of the measures of teachers' perceptions of their work environment had significant direct effects on student progress

As mentioned earlier, most studies within the school effectiveness research literature have been restricted to cross-sectional designs or have made use of pre- and post- data only (see pp. 3ff). Modelling effectiveness in such circumstances is usually accomplished by fitting conditional models in which measures of students background, ability and prior achievement are employed as covariates.
In the case of the VQSP, which has repeated measures on three occasions, namely at the end of the 1992, 1993 and 1994 academic years, it is possible to consider fitting other kinds of models for studying student progress, including, for example, fixed occasion repeated-measures models (Bosker & Scheerens, 1995; Bryk & Raudenbush, 1992, chp. 6; Goldstein, 1995, chp. 6). On the other hand, such models have limitations in the context of the present study. Reference has been made to the problem of sample attrition over the three years of the project. While it is not necessary to have data on all occasions for all subjects to estimate the parameters of a fixed-occasion growth model, with observations on only three occasions, it is not possible to obtain good estimates of intercept and slope coefficients for each student when there is substantial missing data.

Another complication arises from the desire to model the grouping of students within classes and teachers. Most studies routinely take into account the grouping of students within schools, but relatively few are designed to accommodate grouping of students within classes within schools, despite substantial evidence (see Hill & Rowe, 1996) that "...how much a student learns depends on the identity of the classroom to which that student is assigned" (Monk, 1992, p. 320). Similarly, Stringfield (1994) emphasises: "Whatever effects schools have on students, those effects are mediated by teachers" (p. 57). However, those studies that have collected information on class membership and teacher effects have typically been based on cross-sectional designs.

Nevertheless, modelling class grouping effects within a longitudinal study poses special problems because of the constantly changing membership of classes over time. Within a single year, class membership in elementary schools tends to remain relatively intact and the class can be modelled at level-2 within a three- or four-level model. However, there is little constancy in class composition from year to year, at least within the Australian context. Indeed, in many schools, the composition of each class changes beyond recognition as the school attempts to respond to the changing learning needs of students, to accommodate changing enrolments and availability of staff and to implement policies and programs.

Yet another problem, and one which tends to afflict longitudinal studies especially, arises from the situation in which the researchers wish to vary the data-gathering instrumentation during the course of a study. In the present study, as evidence accumulated during the first two years for very large variations in student achievements at the class/teacher-level (see Hill & Rowe, 1996), it was realised that in the initial design of the study insufficient attention had been given to instrumentation designed to measure key aspects of classroom instruction. This led to the development of a set of nine scales referred to above which elicit students' opinions on various aspects of the teaching they experience. These scales were administered in 1994 only.

The above considerations led to the decision to model elementary students' progress in Literacy, for example, using a repeated measures conditional model of the following general form:

\[ y_{ijk} = (X\beta)_{ijk} + \beta x z_{ijk} + (v_i + u_{tjk} + e_{rjk}) \]  

in which the achievement of student \( i \), either at \( t = 1 \), namely at the end of 1993, or at \( t = 2 \), namely at the end of 1994, is predicted by a set of student background characteristic variables \( X_{ijk} \) and a measure of prior achievement \( z_{ijk} \) taken at the end of 1992 when \( t \)
61 Victorian Quality Schools Project

\[ v_{tk}, u_{tjk}, \text{ and } e_{tijk} \text{ are random terms for school } k, \text{ class } j \text{ and student } i, \text{ respectively.} \]

The model described by equation [14] can be represented as a four-level model, with observations at \( t = 1 \) and \( t = 2 \) nested within students within classes/teachers within schools. To allow for the fact that class composition changed from \( t = 1 \) to \( t = 2 \), both 1993 and 1994 classes are modelled as a set of dummy indicator variables at level 3, with estimates of between-class/teacher residual variance constrained to be equal across 1993 classes and also across 1994 classes. Note that no level 1 random terms appear in the model, since this defines the multivariate structure (see Goldstein, 1995, chapter 4 and Rowe & Hill, in press). Thus, in the random part of the model, there are three variables representing student, class and school effects. Note that equation [14] assumes that for each student there are potentially two records or sets of observations, one relating to achievement at the end of 1993 and including a 1992 prior achievement measure, and the other relating to achievement at the end of 1994 and including a 1993 ‘prior achievement’ measure.

In the results reported below for students’ progress in Literacy achievement in elementary schools, a total of 6423 complete records were available for analysis. Of these, 3767 were for student achievement in 1993 and 2656 for achievement in 1994. A total of 1884 of these observations represented students with complete data for both 1993 and 1994. This yielded a total of 4539 unique students.

Table 11 summarises the results of fitting four versions of equation [14] to the data using \( MLn \) (Rasbash and Woodhouse, 1996). Model 1a predicts achievement progress solely on the basis of the Grade/Year level of the student. Models 1a and 1b are essentially equivalent, except that in the case of Model 1a, the estimates for each of the dummy variables representing the random class effect has been constrained to be equal to the estimates for all other classes. In Model 1b, estimates for 1993 classes have been constrained to be equal, as have the estimates for 1994 classes. Thus, Model 1a provides an overall estimate of the residual variance at the class/teacher-level, whereas Model 1b provides separate estimates for the 1993 and 1994 responses.

Considering first the fixed parameter estimates of Model 1b, the variables 1993 and 1994 are the achievement levels (expressed in standardised units and adjusted for all other

---

15 Equation [14] is readily specified in \( MLn \) (Rasbash and Woodhouse, 1996a) using the standard approach to organising data for a bivariate model, with data sorted by school, student and occasion. To indicate class membership, each class was numbered sequentially within each school. The maximum number of classes in any one school was 16 in 1993 and 14 in 1994; thus, the class identification codes ranged from 1 to 16 for 1993 classes and from 17 to 30 for 1994 classes. The final step was to ensure that the appropriate indicator constraint matrix was generated and attached to the model. In the case of Model 1a, each of the variance estimates for all classes at level 3 were constrained to be equal. For Models 1b, 2, 3, 4 and 5, estimates for 1993 responses were constrained to be equal and estimates for 1994 responses were constrained to be equal. Instructions relating to specifying a cross-classified model with the appropriate constraints are contained in Appendix E of the \( MLn \) Command Reference manual (Rasbash & Woodhouse, 1996b, pp. 84-96).

Our preference for \( MLn \) over alternative multilevel analysis software packages (see Kreft, de Leeuw & van der Leeden, 1994) is due principally to its extreme efficiency and interactive flexibility in being able to handle analyses of this kind. Moreover, \( MLn \) is the only currently available package capable of fitting the multivariate multilevel models presented here.

Centre for Applied Educational Research, The University of Melbourne
exploratory variables in the model) of students in the earliest years of schooling, namely Grade 1 and Grade 2, in 1993 and 1994. Then follow dummy variables for the other grade levels, namely Grades 3 to 6, so that the associated coefficients represent the differences from the Grade 1 and Grade 2 values. Model 1b predicts progress scores solely on the basis of Grade level, thus providing a 'null' model with which to evaluate the effects of including further explanatory variables. By adding the coefficients for each Grade level to the coefficients for the base Grades (Grade 1 in 1993 and Grade 2 in 1994), estimates can be obtained of average, gross achievement progress and (from Model 2) of average adjusted, or 'value-added' progress across Grades 1 to 6.

From the random part of Model 1b, it will be noted that the residual variation at the school level is slightly higher for 1994 responses than for 1993 responses. The proportion of variance attributable to between-school differences is about nine per cent for 1993 responses, and 11 per cent for 1994 responses. It will be noted also from the covariance term ($\sigma_{\text{V10}}$) that there is a strong, positive correlation at the school-level between unadjusted 1993 and 1994 achievement scores ($r = 0.842$). The parameter estimates for residual variance at the class/teacher-level (level-3) of Model 1b indicate that there is an intra-class correlation of 0.326 for 1993 responses and 0.371 for 1994 responses, indicating substantial differences among classes in teacher ratings of student achievement. At the student level, estimates of residual variances for 1993 and 1994 are very similar and are highly correlated (0.716).

In Model 2, student Literacy achievement is predicted not only by Grade level, but also by five student background characteristics and by Prior Achievement measured one year earlier. The parameter estimates for Gender indicate that female students make greater progress than males, and students from higher than average Occupational Status families make greater progress than students from families where the highest breadwinner is unemployed or in an unskilled occupation. Having adjusted for other background factors and for Prior Achievement, Non-English Speaking Background, being a student with a Disability/Impairment and having experienced a Critical Event during the year, are associated with significant negative effects on student progress. As expected, Prior Achievement is the most salient predictor of student progress.

Inclusion of the student background variables and the measure of Prior Achievement results in a significant improvement in model fit as indicated by the large reduction in the log likelihood ratio deviance statistic. At the school-level, it also results in a reduction of 24 per cent of the residual variance for 1993 responses, and a reduction of 50 per cent for 1994. At the student-level, there is a reduction of 37 per cent in residual variance for 1993 responses and a reduction of 40 per cent in 1994 responses. On the other hand, there is a small increase in residual variance at the class/teacher-level of nine per cent for 1993 responses and an increase of six per cent for 1994 responses.

In Model 3 the effects of a further seven predictors are estimated. Of these the most influential are student Attentiveness and whether the teacher had attended a specialist literacy in-service course (Literacy PD Program), although the coefficient for the interaction term (Grade x Literacy PD Program) indicates that the effect of participation in such programs was more positive for students in the lower grades than in higher grades. Converting parameter estimates to Effect Sizes (ES) yields estimates for
## Table 11. Parameter Estimates (and Standard Errors) for Four Multivariate Multilevel Models of Students' Literacy Achievement Progress

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Model 1a</th>
<th>Model 1b</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1993 Grade 3 t=1</td>
<td>-1.050 (0.053)</td>
<td>-1.049 (0.052)</td>
<td>-0.521 (0.053)</td>
<td>-0.703 (0.090)</td>
</tr>
<tr>
<td>1994 Grade 4 t=2</td>
<td>0.873 (0.064)</td>
<td>0.874 (0.061)</td>
<td>0.321 (0.063)</td>
<td>0.556 (0.076)</td>
</tr>
<tr>
<td>Gender (Female)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-English Speaking</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disability/Impairment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupational Status</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Critical Events</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior Achievement (1992)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attentiveness</td>
<td></td>
<td></td>
<td>0.044 (0.010)</td>
<td>0.348 (0.010)</td>
</tr>
<tr>
<td>Literacy PD Program</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade x Lit. PD Prog.</td>
<td></td>
<td></td>
<td>-0.179 (0.011)</td>
<td>-0.462 (0.083)</td>
</tr>
<tr>
<td>High Expectations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriateness</td>
<td></td>
<td></td>
<td>0.035 (0.015)</td>
<td>0.006 (0.011)</td>
</tr>
<tr>
<td>Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1993 x Multigrade Class</td>
<td></td>
<td></td>
<td>-0.016 (0.013)</td>
<td>-0.160 (0.056)</td>
</tr>
<tr>
<td><strong>Random:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>School</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2_{v0}$ (1993)</td>
<td>0.044 (0.016)</td>
<td>0.046 (0.016)</td>
<td>0.035 (0.014)</td>
<td>0.032 (0.012)</td>
</tr>
<tr>
<td>$\sigma^2_{v1}$ (1994)</td>
<td>0.064 (0.023)</td>
<td>0.060 (0.023)</td>
<td>0.030 (0.017)</td>
<td>0.034 (0.010)</td>
</tr>
<tr>
<td>$\sigma^2_{v10}$ (1993,1994)</td>
<td>0.044 (0.015)</td>
<td>0.044 (0.015)</td>
<td>0.006 (0.011)</td>
<td>0.009 (0.011)</td>
</tr>
<tr>
<td><strong>Class</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2_{u}$</td>
<td>0.179 (0.013)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2_{u(93)}$ (1993)</td>
<td></td>
<td>0.162 (0.016)</td>
<td>0.177 (0.017)</td>
<td>0.150 (0.014)</td>
</tr>
<tr>
<td>$\sigma^2_{u(94)}$ (1994)</td>
<td></td>
<td>0.202 (0.023)</td>
<td>0.215 (0.024)</td>
<td>0.204 (0.023)</td>
</tr>
<tr>
<td><strong>Student</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2_{e0}$ (1993)</td>
<td>0.289 (0.007)</td>
<td>0.289 (0.007)</td>
<td>0.182 (0.004)</td>
<td>0.145 (0.003)</td>
</tr>
<tr>
<td>$\sigma^2_{e1}$ (1994)</td>
<td>0.283 (0.008)</td>
<td>0.283 (0.008)</td>
<td>0.170 (0.005)</td>
<td>0.133 (0.004)</td>
</tr>
<tr>
<td>$\sigma^2_{e10}$ (1993,1994)</td>
<td>0.205 (0.006)</td>
<td>0.205 (0.006)</td>
<td>0.010 (0.004)</td>
<td>0.008 (0.003)</td>
</tr>
<tr>
<td>$-2^{log(lh)}$</td>
<td>10186</td>
<td>10184</td>
<td>8512</td>
<td>7055</td>
</tr>
<tr>
<td><strong>Intra-school correlations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1993 responses</td>
<td>0.093</td>
<td>0.089</td>
<td>0.097</td>
<td></td>
</tr>
<tr>
<td>1994 responses</td>
<td>0.110</td>
<td>0.072</td>
<td>0.092</td>
<td></td>
</tr>
<tr>
<td><strong>Intra-class correlations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1993 responses</td>
<td>0.326</td>
<td>0.449</td>
<td>0.456</td>
<td></td>
</tr>
<tr>
<td>1994 responses</td>
<td>0.371</td>
<td>0.517</td>
<td>0.550</td>
<td></td>
</tr>
<tr>
<td><strong>Decrease in residual variance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School-level</td>
<td>0.239</td>
<td>0.500</td>
<td>0.304</td>
<td>0.433</td>
</tr>
<tr>
<td>Class/teacher-level</td>
<td>-0.092</td>
<td>-0.059</td>
<td>0.074</td>
<td>-0.010</td>
</tr>
<tr>
<td>Student-level</td>
<td>0.370</td>
<td>0.399</td>
<td>0.491</td>
<td>0.530</td>
</tr>
</tbody>
</table>

Centre for Applied Educational Research, The University of Melbourne
Atteniveness in 1993 of ES = 1.22 and 1.28 in 1994, and for Literacy PD Program of ES = 0.60 and 0.65 in 1993 and 1994, respectively.\(^{16}\)

From an inspection of the estimates in the random part of Models 2 and 3 it is clear that the main effect of adding further explanatory variables is to reduce school- and student-level residual variance, particularly at the student-level. It should be noted, however, that neither Model 2 nor Model 3 results in any significant reduction in the residual variance at the class/teacher-level. Moreover, as more explanatory variables are added, so the correlation among the residual variances at the school- and student-levels reduces markedly (0.186 at the school-level and 0.056 at the student-level).

The models presented in Table 11 are concerned with predicting progress in student literacy achievement. It is also of interest to look at outcomes other than cognitive ones. For example, Model 3 indicates that Atteniveness is an influential predictor of Literacy achievement. But what are the predictors of Atteniveness? Table 12 summarises the results of modelling student Atteniveness as given by Model 4 and Model 5.

From Table 12, Model 4 has no explanatory variables in the fixed part of the model, other than Grade level, but is useful for examining variance components. The estimates for the random part of the model indicates that around 93 per cent of the variance in Atteniveness is at the student-level, with most of the remainder at the class/teacher-level. Between-school differences are either non-existent (for 1993) or trivial (for 1994).

Model 5 incorporates eight further explanatory variables. The strongest predictor of student Atteniveness is Gender, with female students being rated by their teachers as significantly more attentive than their male counterparts. The estimate for the coefficient for the dummy variable for female students of 0.485 yields effect sizes of ES = 0.55 for 1993 and 0.61 for 1994, which, by any criterion, are large effects. Other significant predictors include whether or not the student had experienced some Critical Event during the year (ES = -0.30 and -0.34), the Occupational Status of a student’s parent, student ratings of the adequacy of the Time available to complete tasks in class (ES = 0.16 and 0.18), the Appropriateness of classroom instruction in terms of its level of difficulty, student Incentive or motivation to learn, and teacher Warmth (ES = 0.18 and 0.26). The random part of Model 5 once again indicates that there is negligible residual variation among schools and only a small proportion among classes.

Comments

Consistent with findings at all three data-collection stages of the VQSP (and highlighted in Point 6, pp. 45-46 above), the data presented here again provide strong support for the notion of teacher effectiveness and that it is the identity of the class to which the student belongs that is the key determinant of progress made by the student. However, this finding raises two issues requiring comment. First, it is not possible from these data to estimate the proportion of variance due to between-school differences that may be due to possible unreliability in teachers’ assessments of student performance using the Profiles, since the two are confounded in the design of the study. On the other hand,

\(^{16}\) Following Tymms et al. (1995), effect sizes (\(\Delta\)) for dichotomously scored variables were computed as \(\Delta = \beta/\sigma_e\), and for continuous variables as \(\Delta = 2\beta/\sigma_e\). It should be noted that these Effect Sizes are conditional on other variables in the model.
inter-rater and test-re-test reliability estimates would suggest that the effect of this confounding is not substantial (see Tables 1 and 2, pp. 16-17).

Table 12. Parameter Estimates (and Standard Errors) for Two Models of Student Attentiveness

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1993</td>
<td>-0.071 (0.036)</td>
<td>-0.442 (0.130)</td>
</tr>
<tr>
<td>Grade 3</td>
<td>0.048 (0.052)</td>
<td>-0.370 (0.057)</td>
</tr>
<tr>
<td>Grade 5</td>
<td>-0.049 (0.051)</td>
<td>-0.696 (0.061)</td>
</tr>
<tr>
<td>1994</td>
<td>-0.056 (0.047)</td>
<td>-0.609 (0.130)</td>
</tr>
<tr>
<td>Year 4</td>
<td>0.033 (0.061)</td>
<td>-0.303 (0.065)</td>
</tr>
<tr>
<td>Year 6</td>
<td>0.090 (0.061)</td>
<td>-0.500 (0.070)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td>0.422 (0.023)</td>
</tr>
<tr>
<td>Prior Achievement (1992)</td>
<td></td>
<td>0.395 (0.017)</td>
</tr>
<tr>
<td>Critical Events</td>
<td></td>
<td>-0.212 (0.028)</td>
</tr>
<tr>
<td>Time (1994)</td>
<td></td>
<td>0.123 (0.015)</td>
</tr>
<tr>
<td>Occupational Status</td>
<td></td>
<td>0.113 (0.013)</td>
</tr>
<tr>
<td>Warmth</td>
<td></td>
<td>0.106 (0.021)</td>
</tr>
<tr>
<td>Incentive (1994)</td>
<td></td>
<td>0.096 (0.016)</td>
</tr>
<tr>
<td>Appropriateness of Instruction (1994)</td>
<td></td>
<td>0.060 (0.015)</td>
</tr>
<tr>
<td><strong>Random:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2_{v0}$ (1993)</td>
<td>0.009 (0.007)</td>
<td></td>
</tr>
<tr>
<td>$\sigma^2_{v1}$ (1994)</td>
<td>0.009 (0.008)</td>
<td></td>
</tr>
<tr>
<td>$\sigma^2_{v10}$ (1993,1994)</td>
<td>0.002 (0.005)</td>
<td></td>
</tr>
<tr>
<td>Class</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2_{w0}$ (1993)</td>
<td>0.083 (0.012)</td>
<td></td>
</tr>
<tr>
<td>$\sigma^2_{w1}$ (1994)</td>
<td>0.096 (0.015)</td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2_{p0}$ (1993)</td>
<td>0.664 (0.016)</td>
<td></td>
</tr>
<tr>
<td>$\sigma^2_{p1}$ (1994)</td>
<td>0.558 (0.016)</td>
<td></td>
</tr>
<tr>
<td>$\sigma^2_{p10}$ (1993,1994)</td>
<td>0.285 (0.014)</td>
<td></td>
</tr>
<tr>
<td>$-2*\log(ll)$</td>
<td>16482</td>
<td>15251</td>
</tr>
</tbody>
</table>

The second issue requiring comment is that even from a cursory examination of the findings presented in Table 11, it is clear that the large proportion of variance in students' literacy progress at the class/teacher-level is residual variation that is inadequately accounted for by the fitted explanatory variables at that level, namely Literacy PD Program and Multigrade Class. Moreover, none of the measures of teachers' perceptions of their work environment, including the Leadership Support scale, were found to have significant direct effects on student progress. Nonetheless,
such findings are useful to the extent that they point to a need for a refocus in the prevailing school and educational effectiveness research agenda to one that is closer to students' experiences of schooling in class groupings, and re-examines class/teacher influences on progress in student achievement as advocated by Brophy and Good (Brophy, 1981, 1986; Brophy & Good, 1986; Good & Brophy, 1984), and more recently by Creemers (1992), Slavin (1994, 1996) and Stringfield (1994).

Being in a multigrade or 'composite' class was associated with a substantial negative effect in 1993 but not in 1994. Why was this effect so short-lived? A possible explanation lies in the fact that extended discussions were held with all participating schools following the finding of a negative effect at the end of 1993 and that as a result, schools closely examined teaching practices in multigrade classes with a view to identifying ways in which they have become less effective than single-Grade classes. The finding of no significant effect on Achievement in 1994 is more in line with the research literature at large (see especially the meta-analysis reported by Veenman, 1995).

In terms of the predictors of students' progress in literacy Achievement other than home background factors, prior achievement and the expected negative effects of Non-English Speaking Background, being a student with a Disability/Impairment and having experienced a Critical Event during the year, the two most influential in the positive direction were Attentiveness and whether or not the teacher had participated in a recent, structured literacy professional development program (Literacy PD Program). Interestingly, students' perceptions of teachers having High Expectations of them and ensuring a good match between instructional level and the ability of the student (Appropriateness and Time) were found to be significant predictors, confirming findings from previous studies (see Slavin, 1994). Unfortunately, measures of these constructs were used only in the final year of data collection and thus the estimates relating to these measures relate only to 1994 responses.

Of the predictors of student literacy Achievement, again one of the most salient was students' attentiveness in the classroom (see pp. 47-49), with by far the major proportion of the variance in Attentiveness being evident at the student-level. The most influential positive predictors of Attentiveness were Gender and Prior Achievement, with female students being significantly more attentive than male students, and highlighting the importance of early achievement in affecting students' subsequent Attentiveness. Not surprisingly, Critical Events had a strong negative effect on students' Attentiveness.

The results are also noteworthy for the way in which they clarify the importance of parental Occupational Status and of teacher Warmth as variables that would appear to have significant indirect effects on student progress through their direct effects on Attentiveness. Other positive predictors of Attentiveness include students' Incentive to learn, Time to complete work and Appropriateness of instruction. It will be noted that Time and Appropriateness had both direct and indirect effects on students' literacy progress. These two measures are concerned with whether students feel that they had 'too much' or 'too little' time to complete classwork, and whether the work set is 'too hard' or 'too difficult'. Clearly, students make greatest progress when instruction is pitched at the right level for each individual. These two variables, and of course student Attentiveness, are closely related to the concept of effective learning time, which research has consistently shown leads to positive student learning outcomes (Scheerens, 1992).
14. Schools have considerable influence in overcoming inequalities imposed on students by their background characteristics

The report of the first major empirical study of school effectiveness, *Equality of Educational Opportunity* concluded: "...that schools bring little influence to bear on a child's achievement that is independent of his background and general social context" (Coleman et al., 1966, p. 325). The student-level analyses of the VQSP data can be seen as confirming one of the key conclusions of recent school effectiveness research, namely that the effects of home background are not as pervasive as has typically been assumed.

Certainly, home background characteristics have an impact, for example on secondary students' literacy achievement, but their effects are often not direct since they are mediated through other factors over which schools have a measure of influence, such as attendance at parent/teacher interviews and information evenings (in the case of elementary students) and attentiveness in the classroom. Some characteristics, such as non-English speaking background, which is often assumed to be negative in its impact, have positive effects, for example, on attitudes to school (see Figure 7, p. 51). In addition, effects can be quite complex. For example, it is evident that intake characteristics such as family socio-educational level, gender and rural location exert a greater influence on progress in literacy (mediated via Attentiveness) than on mathematics, indicating that effects can be quite specific within the curriculum.

15. Parents' involvement in monitoring their child's education is important for elementary students

The parent response data (see Hill, et al., 1993a) indicated that there are lower levels of parent participation at the secondary level than at the elementary level and that, at both elementary and secondary levels, parents have generally positive perceptions of their children's school. In modelling the relationships among student-level factors, it was found that at the elementary school level parent involvement in parent/teacher interviews, attendance at information evenings and other activities connected with monitoring their child's progress have a positive effect on student attitudes, behaviour and learning. This finding is significant in that it suggests school intervention strategies that could increase this kind of participation.

16. Schools find data obtained through participation in the project to be useful in the context of school improvement

A major aim of the VQSP has been to facilitate school improvement processes within participating schools. To this end, at the numerous meetings with staff from participating schools, attention was devoted to exploring ways in which each school's results might be used both in the context of traditional approaches to school improvement and more modern approaches to quality assurance and total quality management. In particular, emphasis was placed on exploring the relevance of the information provided as a result of participation in the project to a quality assurance approach that emphasises continuous improvement of processes and outcomes, understanding the reasons for variation, measurement of processes and basing decisions on facts and data, and establishing a client orientation (see Rowe et al., 1993).

Centre for Applied Educational Research, The University of Melbourne
The overwhelming reaction from schools is that they value feedback highly on their own school's pattern of results. In many instances it has provided data of a kind which the school has not had access to in the past, and provided new insights of trends that have been suspected but not confirmed. Staff from the participating schools have also responded positively to the overall findings of the project and have been intensely interested in relating their own results to these findings. The longitudinal nature of the project meant that there were opportunities to monitor the impact of involvement in the project on participating schools and to develop strategies for translating research findings on school effectiveness into school improvement.17

Qualitative Case Study

Purpose

The results obtained from statistical analysis of the quantitative data generated by the VQSP, as outlined above, enabled several generalised models of school and teacher effectiveness to be developed. In 1995, the Centre for Applied Educational Research received an ARC large grant, one aim of which was to conduct a follow-up, qualitative case study of selected VQSP schools in order to illuminate findings from the quantitative modelling of teacher effects on student learning, attitudes and behaviour; and leadership influences on teacher effectiveness. Approval for the conduct of the study was obtained by the relevant human research ethics committees at The University of Melbourne. Specifically, the purposes of the case study were twofold:

1. To 'validate' aspects of the generalised models developed from statistical analyses of the Victorian Quality Schools Project (VQSP) data concerning teacher effects on student learning, attitudes and behaviour; and leadership influences on teacher effectiveness.

2. To illuminate the processes that operate in effective schools (defined in terms of student achievement outcomes) by adding qualitative evidence to the quantitative data already collected.

Research questions

The case study investigations were guided by the following research questions:

1. Do teachers' perceptions of leadership support affect their perception of the nature of the professional community in the school?

2. Do teachers' perceptions of the nature of the professional community in the school affect their perception of the teacher-student interaction in the school?

3. Do leaders' perceptions of their role affect:
   • the practices leaders adopt to fulfil their perceived role?
   • the view teachers have of the support (Leadership Support) given to them by school leaders?

17 For specific details of these strategies, see Rowe et al. (1993).
4. To what extent does teachers’ participation in literacy/numeracy professional development programs have perceived positive effects on students’ progress in English and Mathematics?

5. Do students’ perceptions of teacher responsiveness affect their attitudes to school?

6. Does the predominant use by teachers of certain forms of class organisation and teaching/learning strategies have a positive effect on student progress in Literacy and Mathematics?

7. Does class composition based on more than one Year level have a negative effect on student progress in Literacy and Mathematics?

8. To what extent does differentiated teaching reduce the negative effect on student progress in Literacy and Mathematics of belonging to a class composed of students at more than one Year level?

9. What do teachers believe to be the factors accounting for the plateau in students’ development in Literacy in the upper primary school and lower secondary school?

Sample and methods

Schools

A sample of six primary schools was selected from among those that had participated in the VQSP. The qualitative study was confined to primary schools for two reasons: first, some of the most interesting and important findings of the VQSP related to the primary school and, secondly, time/cost demands of the case study approach precluded the investigation of a sample large enough to include both primary and secondary schools.

Selection of the six schools was based on schools’ mean ‘value-added’ learning progress scores in English and Mathematics for the years 1992 to 1993 and 1993 to 1994. Two schools were selected which had consistently high mean scores, two with consistently low mean scores, and one with consistently middle-level mean scores. The Chief Investigators carried out the selection process, and the Case Study Coordinator and Fieldworkers remained blind to previous performance of the schools, so that their interactions with the school and interpretation of the data would not be influenced. Replacement of one school in the sample was necessary because the Principal in 1995 was new and had not been the school leader for the duration the VQSP; it was essential to have continuity of school leadership for the case study.

All six schools, when approached, agreed to participate in the study. The sample comprised schools from two systems (Government and Catholic), from a range of locations (urban, outer urban and semi-rural), and schools ranging in size from small (125 children, 8 staff) to large (525 children, 27 staff). An assurance of confidentiality was given to the schools and to the individuals who participated.

Students

Students in Grades 3 and 5 were selected for the student sample within schools. Because these were the children who had participated in the VQSP from Prep and Year 2 respectively, their 1995 results could be linked to previous results. Two class groups
from each Year level formed the student sample; in one school there was only one class group at Year 5. The total number of students in the sample was 562.

A small group of 5 to 7 students was selected from each of the two Year levels (with parental consent), representing a range of abilities, to form focus groups for interviews.

**Staff**

The staff sample consisted of:

- four school leaders - the Principal, the Assistant Principal/First Assistant, and the two staff members holding the next most senior positions; in one school there was no Assistant Principal (n = 23);
- the four (in one school three) teachers whose Year 3 and Year 5 classes formed the student sample (n = 23);
- four other teachers from the staff, selected randomly (n = 24);
- all members of staff (n = 98).

**Instruments: Interview schedules**

Semi-structured interview schedules were designed for leaders, class teachers, other teachers and students. All interviews were tape-recorded. Areas covered in the interviews are as follows.

1. **Leaders:** Current challenges for the school; school priorities and effects on these of challenges; important teacher qualities; means of improving the quality of teaching; ways to help maximise student learning; policy about and teaching strategies for composite classes; perception of leadership practices; changes in role of school leaders; factors accounting for the plateau in students' development in English in the upper primary school. These were completed by four leaders in each school.

2. **Class teachers:** Expectations of current class; strategy for planning and teaching a new concept or skill; strategies for teaching and learning in a composite class compared with a single Year level class; relative ease of learning in composite classes; school policy about composite classes; school organisation and maximisation of time on task; literacy and numeracy professional development and its effects; factors accounting for the plateau in students' Literacy development in the upper primary school. These were completed by the four class teachers in each school.

3. **Other teachers:** Current issues for school; effects of a positive professional culture in the school; staff involvement in professional development and the effects; staff goal congruence and its effects; peer support and its effects; leadership support and its effects; perception of leadership practices; factors accounting for the plateau in students' development in Literacy in the upper primary school. These were completed by the four other teachers in each school.

4. **Students:** Good teachers - what they are like and what they do; preferences regarding class organisation for learning; types of teacher help; things children like

---

18 For specific details of all instruments used for the case studies, see Appendix “E”.

Centre for Applied Educational Research, The University of Melbourne
about the school and things they would change. These were completed by the two Year level focus groups of students in each school.

**Instruments: Questionnaires**

1. **Leaders:** Leadership style values; constraints and opportunities; self-concept; attributions of responsibility; sources of information. These were completed by the four leaders in each school.

2. **Teachers:** Literacy and numeracy professional development; self-perceptions; work attitudes and strain; work environment. Relevant instruments were completed by all teachers in each school.

3. **Students:** Homework; quality of school life. Instruments were completed by all students in the target Year 3 and Year 5 classes in each school.

**Other instruments**

1. Ratings of each student's levels of development in English and Mathematics by each of the target Year 3 and Year 5 class teachers, using the Victorian Profiles.

2. Ratings on the RBRI *Attentiveness* scale of each student by each of the target Year 3 and Year 5 class teachers.

3. Specimen record and ratings of Year 3 and Year 5 class teachers on 11 dimensions of teaching/learning strategies and class organisation, during observations of an English and a Mathematics period in each group.

4. Professional diary in which each professional interaction during one day was recorded in terms of topic, duration, format, initiator, and number involved. These were completed by all leaders and teachers in each school.

5. School documents (such as school charters), including other incidental information and observations were collected from each school.

**Procedure**

The Coordinator and Fieldworkers were responsible for data collection in two schools each. The selection of schools ensured that no investigator was responsible for data collection in a school with which they had previously had professional contact.

Following pilot work on some of the instruments in an outer urban primary school, the fieldworkers made a preliminary visit to the first of their two schools in order to establish rapport and make organisational arrangements for data collection. Over a two-week period, fieldworkers spent four to five days (different days of the week) in the school, interviewing, observing and collecting data. The same procedure was followed with the second school. The fieldwork was completed by early November, 1995.

**Data analyses**

Quantitative data were key-punched and analysed using the SPSS and LISREL programs. The initial set of analyses undertaken were those related to student achievement in *Literacy* and *Mathematics*, teacher ratings of student *Attentiveness* in the
classroom, students' perceptions of homework and the quality of school life, teacher ratings of work attitudes and strain, their work environment, and affective self-perceptions. A summary of their results was sent to each school in February, for early feedback on the project and for use in school annual reports.

Interview responses were transcribed (not verbatim) from the tape-recordings and, following the methodology of Miles and Hubermann (1994), were used to establish within-site matrices relating to each research question and subsequently across-site matrices.

Results

Given the richness of the qualitative data obtained, together with the amount of quantitative data collected, it is not surprising that the results of the study were extensive as well as detailed. The data relating to each research question were considered in terms of the two aims of the study, namely: (1) whether they validated the generalised models developed from the VQSP, and (2) whether they illuminated the processes proposed in these models. Both aims of the study were achieved.

For illustrative purposes, several examples are provided here. An example of the range of analyses undertaken in relation to one of the research questions (Research Question 5) and the results of those analyses is provided in Appendix 1 of this section.

One finding from fitting the explanatory models of student progress in primary English and Mathematics was the consistently strong influence that Teacher Responsiveness exerted on students' Attitude to School (see Figure 7, p. 51). The case study data provided both quantitative and verbal data supporting this influence. Student interview data also broadened the construct of Teacher Responsiveness beyond its focus on provision of help with learning when the student is having difficulty, to include helpfulness in many other aspects of the management of learning inside the classroom, together with helpfulness outside the classroom. Other dimensions of teacher behaviour and personality were also perceived by students to be important: individual interest in students; fairness, caring about students' well-being, and treating students with respect and friendliness.

Analyses of the 1992 and 1993 data indicated a significant negative effect on student learning progress from membership in a class comprising more than one year level (see Figure 7, p. 51 and Table 11, p. 63). However, analyses of the 1994 data showed no significant difference on the basis of class composition, although the trend was still negative (see p. 66). The comparison of student learning progress in Literacy and Mathematics in the case study sample showed no significant differences in student progress in 1995 between students in single year classes and those in composite classes. While this was contrary to the findings of the study in 1992 and 1993, it supported the 1994 findings.

Teacher and leader perceptions of appropriate teaching/learning strategies for composite classes, of the relative difficulty of learning in composites as opposed to single year level

---

classes, and of the most appropriate student selection for such classes, were obtained in interviews. In contrast to teacher responses about perceived Leadership Support and its effects or the perceived effects of a positive professional culture in a school on teacher-student interaction which were consistent across teachers and schools, perceptions about teaching and learning in composite classes constituted a range of contradictory views about each aspect explored. There appeared to be no generally accepted body of professional knowledge or practice in schools relating to this question; instead, there appeared to be something approaching a collection of popular myths. If this is the case, it is not surprising that performance of students in composite classes is variable. Nonetheless, this area requires further investigation.

Findings from fitting the explanatory models of student progress and of school leadership indicated a strong influence of Leadership Support on the Professional Culture of the staff within the school (see Figure 9, p. 55 and Table 9, p. 56). Case study interview data confirmed the influence of the constituent aspects of Leadership Support included in that survey scale. Teachers perceive leader approachability, understanding, communication and back-up to be important aspects of supportiveness.

Teacher comments also enriched the understanding of the Leadership Support construct by pointing to the fact that it is not only what leaders do that matters, but the manner in which these things are done. For example the critical feature of a leader’s willingness to engage in conversation about a teacher’s concerns is not simply that the interaction takes place, but that the leader should be genuine in his/her listening and interest. The interview data also extended and elaborated the construct of Leadership Support. Other important dimensions for teachers are fairness, individualised support, modelling good practice, and professional stimulation. This enriched and extended concept of Leadership Support could well be included in the development of additional survey instruments and used for diagnostic purposes in schools.

The student-level results indicated a plateau in their English development in the upper primary school (see Figure 1.1, p. 32). Teacher and leader perceptions of the reasons for this plateau were explored. Suggested explanations fell into three categories.

- **Transition**: For example, secondary school size and complexity; change from most senior to most junior group; repetition in curriculum; information gaps; differences in the teaching strategies/orientations of Year 6 and Year 7 teachers.

- **Developmental stage**: For example, intellectual consolidation; Year 6 behaviour and attitudes; Year 6 student responsibilities and activities; peer pressure; hormones; attitudes to learning; decline in parental involvement; embarrassment of those not coping.

- **Curriculum**: For example, over-crowded curriculum; repetitive, unchallenging curriculum; preparation for secondary school; unrelatedness to children’s specific interests; the nature of the English (Literacy) curriculum compared with that for Mathematics.

These suggested explanations require follow-up in future research. There is considerable concern in schools about the need to improve education for students in the middle years (Years 5 to 8). Because of this, some schools are already examining the implications of
these findings in their efforts to bring about the development of improved middle school programs for students.

Conclusion

The case study on school and teacher effectiveness met the two constituent aims which were to 'validate' aspects of the generalised models developed from the statistical analyses of the data, and to illuminate the processes that operate in effective schools.

Dissemination of findings from the study is on-going, through presentation of papers to conferences of Principals and teachers, and through the provision of specific feedback to participating schools. Findings continue to be disseminated through paper presentations at academic conferences and publications in journal and monograph form.

APPENDIX 1. RESEARCH QUESTION 5: ANALYSES AND RESULTS

Do students' perceptions of teacher responsiveness affect their attitudes to school?

SOURCE OF HYPOTHESIS

The explanatory model of student progress in primary English and Mathematics based on two levels of data (teacher and student) indicated that Teacher Responsiveness had a strong and significant positive effect on students' Attitude to School. The relevant scale definitions and response items are as follows:

**Teacher Responsiveness.** Items in this scale are designed to indicate the extent to which students perceive that their teacher is helpful and responsive to their learning needs. *Normative scale statistics, reliability, response items and item factor score regression weights* (M = 3.37; SD = 0.53; SEM = 0.01; r_e = 0.824):

4. My teacher helps me with my work (.276).
13. My teacher shows me how to do things when I am having difficulties (.403)
22. My teacher cares if I am not doing as well as I should (.091)
31. My teacher takes time to help me when I have trouble with my work (.230).

**Attitudes to School - Enjoyment.** High scores on this scale are indicative of students who enjoy being at school and express enjoyment of their school-based learning experiences. *Normative scale statistics, reliability, response items and item factor score regression weights* (M = 3.09; SD = 0.60; SEM = 0.01; r_e = 0.866):

1. I enjoy the work I do at school (.264)
5. Learning is fun at my school (.362)
9. I get excited about the work I do at school (.209)
13. I enjoy being at my school (.165).

SUMMARY OF FINDINGS

1. Within the case study sample, students' Attitude to School (Enjoyment) and their ratings of Social Acceptance, Motivation and perceived Usefulness of the curriculum, were far from uniform. There were:

* wide differences among individual students;
• significant differences among teacher/class groups, both within and between schools;
• significant differences between schools;
• significant differences between Years 3 and 5 students.

These findings were consistent with the quantitative results (see, for example, Figure 4, p. 38) and provided validation support for the large positive effect of teachers' influence on students' attitudes to school obtained from fitting the explanatory model. Data from the case study indicated there was greater variability in students' ratings/perceptions of Enjoyment and Social Acceptance than for Motivation and perceived curriculum Usefulness. The greater uniformity in individual students' scores on the Motivation and Usefulness scales were also evident when comparisons were made between teacher/class groups and between schools.

2. Highly significant differences were found among class/teacher groups on all four scales (with the greatest difference being on the Enjoyment scale), indicating the importance of specific teachers to the nature of student attitudes. Within four of the six schools there were significant differences between teacher/class groups in student attitudes, especially for Enjoyment and Social Acceptance, indicating the greater vulnerability of these attitudes to specific teacher effects. Again, these findings were consistent with results from the explanatory modelling.

3. Highly significant differences between schools were found for students' attitudes, namely, Enjoyment, Motivation and Usefulness, and to a lesser extent for Social Acceptance. This is to be expected given the selection of sample schools on the basis of two extreme groups and one middle-level group. The differences were suggestive of a school culture effect, especially in the areas of student motivation and perception of the usefulness of learning. It might also have direct links in to the teacher level data, namely the Professional Culture and Teacher-Student Interaction dimensions. It would be very interesting to fit a further explanatory model tracing the influences on student affect rather than on their learning progress.

4. Highly significant correlations between the four quality of school life dimensions suggested an underlying general factor of positive affect towards school and schooling. However, discrimination by students among the four dimensions indicated that the four scales tapped different aspects of such an underlying dimension. Student comments reinforced these points; so too did the size of the squared correlation coefficients. An explanatory model focussed on affective outcomes would elaborate on this.

5. In general, Year 3 students had significantly more positive attitudes to school than did Year 5 students: they enjoyed school more, felt more socially accepted and were more highly motivated. The one attitude that did not differentiate between the age groups was that of perceived Usefulness, with the observed means for the two groups being almost identical. While the data demonstrated declining attitudes in student enjoyment (see Figure 4, p. 38), it is interesting to note that students' perception of Usefulness was not less strong in Year 5.

6. Essentially, the Teacher Responsiveness scale focused on the extent to which students perceived that teachers helped them with work, especially when having difficulty. Student comments validated this as an important aspect of teacher helpfulness.
Student comments also extended and illuminated the concept of teacher helpfulness when students have difficulties, for example good teachers:

- have established procedures to try to ensure students indicate if having trouble
- are expert at noticing, interpreting body language and reading signs, if students don’t tell them
- give help patiently, without yelling and getting angry if child doesn’t understand
- persevere, if a child doesn’t understand, showing in different ways, keeping on trying, and wanting the child to understand.

7. Students saw the helpfulness of the teacher as extending beyond occasions when they have difficulty in learning to:

- other aspects of the management of learning, for example:
  - ability to explain clearly when learning new things
  - use of stimulating and enjoyable activities
  - making the level and pace of learning appropriate to the students
  - spending appropriate periods of time on work, with breaks in between
  - achieving good learning outcomes
  - motivating student, giving them confidence and reassurance
  - giving positive feedback and rewards
  - maintaining a quiet classroom environment in which they can learn.

- other aspects of school life, for example:
  - alertness to possible dangers in the yard, warnings about syringes etc.
  - ministering to them when they are hurt (*Band Aids are BIG in their lives!*).

8. Evidence of other aspects of teacher responsiveness that were important to student attitudes was seen in student comments. These aspects focused on the way in which the teacher related to the student; some of these dimensions were tapped in the student questionnaire *Opinions about Teachers and Teaching*, others are not. Examples were:

- individual interest in students - knowing names, talking to students, being interested, keeping confidences, taking them seriously
- manner of interactions - talking nicely (“nice” is a favourite word), being kind, caring, considerate, respecting students, friendly
- fairness in various ways - not playing favourites, not picking on someone, delaying blame until evidence considered, making sure everyone is listening and learning
- caring about students’ well-being - working hard to teach well, keeps safe and comfortable, solves problems.

These aspects confirm the importance of dimensions that were tapped in the student questionnaire *Opinions about Teachers and Teaching*. Student comments provided much colour and detail to our understanding of them.
9. The effects of Teacher Responsiveness, as related by students, again confirmed and extended the findings from the explanatory modelling. The main effects were:

- positive affect – feeling excited about work, being happy to come to school, feeling comfortable in the classroom, feeling safe in the school where there is no violence, the rules are fair
- better behaviour - children are quiet, attentive and cooperative, obey the rules
- better learning - work is completed and on time, understand more, learn faster, achieve standards
- relate well to teacher - nice to them and do nice things for them, respectful, polite, share jokes and laugh with them, not at them
- children are friendly to one another rather than nasty.

10. The influence of Teacher Responsiveness on Attitude to School, and of Attitude to School on Attentiveness from the explanatory modelling of student progress in primary English and maths, was also reflected in these student comments. The comments also suggested ‘cause-effect’ relationships which could be included in an explanatory model focussed on affective outcomes.

11. Information from student responses contained in points 7-9 above could be of use in the development of further instruments and further research.

12. Children at both Year levels were seen to be:

- very astute observers and perceivers of teachers and themselves;
- sensitive responders to adult behaviour and attitudes;
- able to articulate their experiences of human interaction.

CASE STUDY DATA ANALYSES

1.0 Are there differences among students in Attitudes to School?

1.1 Individual differences: YES

- **Enjoyment:** mean = 0.65; median = 0.67; mode = 0.67
  range: +2 to -2; SD = 0.80
- **Social Acceptance:** mean = 0.92; median = 1.09; mode = 0.67
  range: +2 to -2; SD = 0.80
- **Motivation:** mean = 1.54; median = 1.69; mode = 2.00
  range: +2 to -0.91; SD = 0.50
- **Curriculum Usefulness:** mean = 1.52; median = 1.72; mode = 2.00
  range: +2 to -2; SD = 0.63.

1.2 Teacher differences: YES

- **Enjoyment:** \( F = 4.48; p < 0.001 \)
- **Social Acceptance:** \( F = 2.36; p < 0.001 \)
• Motivation: \( (F = 2.83; p < 0.000) \)
• Curriculum Usefulness: \( (F = 2.44; p < 0.000) \).

1.3 Differences among teachers within schools: SOME

• More and larger differences were found on the Enjoyment scale, indicating this dimension was more responsive to the nature of the specific teacher than the others. The Social Acceptance scale also showed significant differences in half the schools, suggesting this dimension was also teacher-related.

• In only one school was there a difference among teachers on the Motivation scale and none on the Usefulness scale. However, since there were significant differences among schools on these dimensions, the school culture appeared to be of greater importance than the individual teacher in these areas.

1.4 School differences: YES

• Enjoyment: \( (F = 6.16; p < 0.001) \)
• Social Acceptance: \( (F = 2.11; p = 0.047) \)
• Motivation: \( (F = 3.26; p = 0.007) \)
• Curriculum Usefulness: \( (F = 5.46; p < 0.001) \).

(Note that the rank order of schools in terms of mean scores on these scales was similar to that of the rank order in terms of student learning progress. The exceptions were the two small schools. These schools seemed to give results which were less predictable in other areas also.)

1.5 Relationships among attitude dimensions

• Enjoyment correlated moderately with Motivation \( (r = 0.44) \), Usefulness \( (r = 0.39) \) and Social Acceptance \( (r = 0.35) \).

• Social Acceptance also correlated with Motivation \( (r = 0.24) \) and Usefulness \( (r = 0.27) \), with smaller slightly smaller correlation coefficients.

• While other results (eg., 1.3 above) showed that students discriminate among the items contributing to the different scales, there may well be a general factor of positive feelings towards school underlying these results.

1.6 Year level differences: YES

• Enjoyment: \( (t = 2.90; p = 0.004) \)
• Social Acceptance: \( (t = 2.11; p = 0.035) \)
• Motivation: \( (t = 2.58; p = 0.010) \)
• Curriculum Usefulness: \( (t = 0.00; p = 0.996) \).

1.7 Relationship between attitude dimensions and rating by observers of teachers' provision of individual attention and help

• A relationship between these variables might be seen to provide a reality basis for student perceptions.

Centre for Applied Educational Research, The University of Melbourne
• Specific relationships were not discernible. Possible reasons for this include: long
term relationship between teacher and students compared with the small number of
observations, absence of observers’ knowledge of students, nature of observation and
its effect on teachers and students, difference between adult observations and student
perceptions, effect of interaction of many teacher dimensions rather than a single
dimension.

2.0 Student perceptions: What characterises teacher responsiveness for students?

2.1 ‘Good teachers’

1. How teachers treat us
- Relating to children
Can tell by their attitude; nice, kind, caring, helpful, understanding, have respect,
polite, considerate, friendly, makes you feel comfortable.
Can tell by way they greet you, talk to you, talks nicely, doesn’t yell.

Interested in you, listens, takes you seriously, talks with you about your problems,
won’t break a confidence, will have friendly chat, talk to you in the school yard and
join in playing games.

Helpful all the time, especially if someone has an accident (sickbay, band aid, ice
pack), watches and cares, always on the alert, will help work it out if someone is
 teasing.

Treat you fairly, don’t play favourites, or pick on one person, take time to work out
things fairly if problem arises. For example:

"Say there’s a problem in the yard, they figure it out. They don’t just yell at
people straight away. Some teachers just say, “You did it.” They don’t bother
to figure it out.”

"Some teachers, if your work’s good they don’t say anything about it. They just
send you back to your desk. Some teachers, say one person did good and some
other person did about as good as them and they don’t even say you did good to
the other person - like they favour certain people."

- Teaching and learning
Breaks for relaxation, not unrelieved work: Make you work hard but not all day,
have some free time, breaks to read stories or play game or do some fun things.

Stimulation: Have interesting subjects and activities, interesting work that suits each
person, lots of different things, fun activities.

Appropriate level and pace: Gives you time to finish your work, gives you hard
work but not too hard, doesn’t go too quickly. For example:

"You have to use your brain and think about it so much that it gets tired, which is
just about the right level for us”

Helps if you’re having trouble with your work, helps rather than yells if you get
something wrong; don’t just explain once and refuse to explain again if you forget
later.
Doesn’t let you muck around, is strict but not too strict, class is better controlled, keeps cool and calm, tries to distract children, reasons with them, gives warnings.

Learn things better and faster, explains things well, clearly, properly, helps you get up to a standard, makes sure you understand everything.

Gives you rewards, makes good comments on your work.

2. How we respond

• **Feelings**

Feel excited, happy, comfortable, not as tense, happy to come to school. (Contrast: Get bored, don’t care, shy and not answer questions, run away from school.)

• **Relating to teacher**

Nice to them, do nice things for them, get others to be nice to them, be respectful, polite, share jokes and laugh. (Contrast: Say things behind their back, call them names, answer back, disobey, inconsiderate, mean, try and get attention.)

• **Learning**

Do the work, finish things on time, learn better, not muck around or call out, obey the rules. (Contrast: Muck around and behave badly, steal chalk, talk, call out, don’t put up hands, run and throw things, swear, take others’ things, don’t obey the rules, act kinda silly, like smart alecs because you don’t care; don’t do work, don’t work as hard, don’t get work finished, don’t do homework.)

3. How teachers make things interesting

• **Subject areas**

Have enjoyable subjects - art, technology, science, drama, craft, cooking, games, write stories. (Contrast: Subjects you don’t like - maths, hard maths, mental maths, writing, language, spelling, LOTE, silent reading.)

• **Stimulating activities and relationships**

Gives us “good” work, something we enjoy, gives us different types of work, uses games in teaching, project work, assignments that use concrete materials, special events and excursions.

Teacher uses different voices, expressions actions in reading (“Your mind gets this funny feeling that you’re really there.”); makes work fun (“Teachers study hard to find things that are interesting for children to work on.”); always smiles, makes jokes; speaks clearly, encourages everyone; gives us work that’s not too hard.

Gives you praise and rewards for good work. (Contrast: Repetition, if finish early going back and doing more, doing things “over and over again”, same subject, same sheets, same work; keeps explaining when you already know what to do; copying down heaps of writing from board; doing work all the time without a break or free time; work that’s too easy or work you don’t understand; work that isn’t explained.)
2.20 Teacher responsiveness when learning new things

2.21. Types of help a teacher gives

- **Level and pace of initial teaching**
  Explains clearly, properly, carefully, slowly.

- **Gives help to those who don’t understand immediately**
  Goes through again with them, several times if necessary, shows how to work out without telling answer, knows who’s likely to get stuck, asks if you understand so can come and get help.

- **Motivation, stimulation, self-esteem**
  Gives you confidence, makes it sound exciting, makes sure you listen, uses varied teaching strategies.

3.0 Whether it matters to a teacher if we understand

- **Keeps trying** to make you learn it, won’t give up, if don’t succeed they think there’s got to be something else they can do, they want you to understand.

- **Cares about you** and want you to learn, they like children, they look happy because you’re trying.

- **Wants to do job well**, they feel they’re not doing their job properly if you don’t catch on.

- **Understands consequences** for children of not learning: won’t get very far, be ready for high school, help with future work and business.
  
  *(Note: Students also perceived consequences for teachers, eg., get into strife from parents; get fired!)*

4.0 How we know if a teacher cares

- **Gives positive feedback**: praise, compliments, good marks, sweets.

- **Listens to you**, listens to your problems, helps sort out your problems in the yard.

- **Talks to you** a lot:
  
  'Says “S- come over here and have a chat” - friendly - that’s my experience here with Mr X -and then he questions me about my work’

  "I do my somersault down the hill and then if the teacher is there he will come up to you and ask ‘How are you?’"

- **Treats you well**, nicely; you can tell by the words they use, her voice, her expression, shows you respect, treats you fairly, “She lets you be what you are.”

- **Helps you when you are hurt**, Band Aids, sick bay, etc; warns you about syringes on oval and tells you not to roll on ground!

- **Perseveres patiently to help you understand**: keeps on helping and showing till you get it right, never just leaves you; won’t say “Go away and figure it out.”;
won't get angry with you for not understanding; if you’re not listening will stop so you’ll learn more; explains and shows how really well.

<table>
<thead>
<tr>
<th>5.0 Whether a teacher notices if we have trouble:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• <strong>Direct approach:</strong> they ask and you tell them.</td>
</tr>
<tr>
<td>• <strong>Established procedures:</strong> put up hand; teacher walks around looking at work.</td>
</tr>
<tr>
<td>• <strong>Reading the signs:</strong> teacher ‘just sees it’; sees signs - dropped pen, the look on your face, not working and just sitting or fiddling. For example:</td>
</tr>
<tr>
<td><em>“Sometimes you can see by the expression on their face that some people don’t really know what to do.”</em></td>
</tr>
<tr>
<td><em>“Sometimes she doesn’t notice if you’ve got your head down; if she did, she’d stop and help you.”</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6.0 What a teacher does when we’re right</th>
</tr>
</thead>
<tbody>
<tr>
<td>• <strong>Correction:</strong> ticks, tells answer.</td>
</tr>
<tr>
<td>• <strong>Verbal feedback:</strong> praise, good comments, ‘Good work!’, ‘Excellent!’, ‘Well done!’, ‘Fantastic!’</td>
</tr>
<tr>
<td><strong>Rewards:</strong> stickers, stamps, smiley, marbles in jar, house/table points, lollies, awards, shows the principal.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>7.0 What a teacher does when we make a mistake</th>
</tr>
</thead>
<tbody>
<tr>
<td>• <strong>Verbal indication and reassurance:</strong> ‘Sorry, you’re wrong’, ‘Doesn’t matter, try again’, says you’ll probably do it next time.</td>
</tr>
<tr>
<td>• <strong>Teacher action:</strong> shows how to work out but doesn’t give answer; explains it better; doesn’t get angry, just shows you how and tells you what doing wrong.</td>
</tr>
<tr>
<td>• <strong>Action required of child:</strong> gets us to re-write the word; after showing how, says rub out and do again; asks you to fix it up. For example:</td>
</tr>
<tr>
<td><em>“I’m not a very good speller. If we have 16 words and I get 2 out of 16 of them right, and say we have to write out the words 10 times, I’m lucky because I only have to write them out 5 times each.”</em></td>
</tr>
</tbody>
</table>
Concluding Comments

We shall not cease from exploration
And the end of all our exploring
Will be to arrive where we started
And know the place for the first time.

(T.S. Eliot, Little Gidding)

The Victorian Quality Schools Project (VQSP) has generated one of the most comprehensive and significant data bases on school and teacher effectiveness ever assembled. It is the first major empirical study of school effectiveness undertaken within Australia and one of few internationally to collect longitudinal measures of schools, teachers and students using a design that has enabled the estimation of effects at the student-level, class/teacher-level and school-level. Analyses of both the quantitative and qualitative data suggest that conventional views on school-level effects need to be re-evaluated in the light of findings from the VQSP. A summary of the key findings have been outlined above (pp. 31-82), from which several noteworthy conclusions can be drawn.

First, to the extent that it is meaningful to talk about educational effectiveness in an Australian school setting, it would appear necessary to acknowledge that explanations will change significantly depending on the way in which criterion measures of 'effectiveness' are operationalised and whether the focus is upon elementary or secondary school students. The results summarised in Table 4 (p. 43) and Table 5 (p. 44) suggests that different kinds of adjustments can lead to different explanations of educational effectiveness and point to the desirability of working with a variety of student progress measures.

The results given in Table 6 (p. 46) indicate that different sets of variables may be significant in explaining student progress in Literacy as opposed to Mathematics, and for elementary school students as opposed to secondary school students. This adds a new dimension to the phenomenon identified by Reynolds et al. (1994) who cite evidence against the proposition of 'across the board' effectiveness. Not only are correlations between measures of effectiveness across different subject areas likely to be modest, indicating differential effectiveness across curriculum areas, but explanations of effectiveness also may vary from subject to subject within those curriculum areas (see Luyten, 1994; Tymms, 1993).

Second, when curriculum-specific teacher assessments are used to assess student achievement, the variables that best explain student progress tend to be those which directly relate to classroom-level instruction. The findings summarised in Table 5 (p. 44) indicate that when a measure of elementary student progress in Literacy is used with adjustments for Grade Level and Prior Achievement, the most important explanatory variables are student Attentiveness in the classroom and teacher participation in specialist Literacy In-Service Programs. These variables measure important features of the classroom teaching and learning environment. By comparison with these classroom-related variables, the effect of student background variables such as Socio-Educational Level, Gender and Non-English Speaking Background tend to be small or mediated by Attentiveness. Indeed, there are clear indications that Attentiveness operates as a suppressor variable in the presence of both Socio-Educational Level and Gender (see Centre for Applied Educational Research, The University of Melbourne
Rowe & Rowe, 1992b). On the other hand, the effects of background variables are more pronounced when a measure of student progress is used with adjustments for Student Ability rather than with Prior Achievement. This suggests that the effect of using highly generalised, standardized achievement tests may be to exaggerate the importance of such background variables, since such measures will not be sufficiently sensitive to detect differential progress within the classroom.

Third, the exploratory modelling summarised in Table 7 (p. 50) and Figure 7 (p. 51) further illustrates the fact that a proper understanding of educational effectiveness is likely to require models that go beyond simple input/process-output models in order to better model the complexity of schools, classrooms and the factors that influence what transpires inside them. This underscores the importance of developmental work on methods for undertaking multilevel analyses of complex structural relationships between networks of variables (for recent developments, see Bryk & Raudenbush, 1992; Goldstein, 1995; Longford, 1993; McDonald, 1994).

Fourth, irrespective of the kinds of adjustments made, there is consistently greater variation in student achievement at the class/teacher level that at the school level. Herein lies a paradox of considerable interest and importance for the understanding of school effectiveness, both in the Australian context and elsewhere. On the one hand, data from the VQSP reveal substantial between-school variance in terms of the socio-educational status of students, and at the secondary level, in terms of student aptitude as measured by standardized tests of verbal and quantitative ability. However, when adjustments are made for intake characteristics, student ability and prior achievement, this between-school variation largely disappears (3% to 8%) and is replaced by very large proportions of between-class/teacher variance, ranging from 38% for secondary students' Literacy achievement to more than 55% for elementary students' achievement in Mathematics.20

Clearly, this finding suggests that a given school is likely to be only as effective as the quality of classroom teaching within that school. That is, the large proportion of between-class/teacher variation is attributable to actual differences in the amount of progress made by individual students within particular classes, reflecting differences in the quality of teaching and students' willingness and ability to profit from instruction within that class. Such an explanation fits with findings from a study in Victorian elementary schools by Ainley, Goldman and Reid (1990) who found that differences among teachers within schools were greater than differences between schools in the growth they achieved in their students. It also fits with a recent Australian national survey of community views of what makes an effective school by McGaw et al. (1992) which found that the most frequently mentioned factor was the quality of the teachers, constituting 65 per cent of all responses. Curiously, it is only relatively recently that school effectiveness researchers have begun to recognise the importance of instructional effects at the class/teacher level on students' achievement outcomes (eg., Creemers, 1992, 1994b; Schaffer, Nesselrod, & Stringfield, 1994;Teddlie, 1994).21

---

20 For specific details, see Table 4, p. 43.

21 A further reason for this has been noted by Gray (1989), Creemers and Reezigt (1996) and by Hill and Rowe (1996), namely that school effectiveness research employing multilevel analysis has been dominated by studies employing two-level models where the predominant analytic structure of the data has been students (level 1) nested within schools (level 2). Relatively few studies have been undertaken which account for class-level effects, despite mounting evidence for doing so.
An alternative explanation for the degree of between-classes variation is a reflection of particular adjustments made to the criterion measure. By adjusting each student's achievement at the end of the year by his or her Prior Achievement at the end of the previous year, it could be anticipated that between-school differences would be small and that the between-classes variance be accentuated. Such an explanation seems unlikely, however, given that the proportion of between-classes variance was even higher when achievement measures were adjusted for Student Ability (see Table 4).

A further explanation for this outcome is that the high levels of between-classes variation reflect unreliability in the measures of student progress. Unfortunately, it is not possible from these data to estimate the proportion of variance due to between-school differences that may be due to possible unreliability in teachers' assessments of student performance using the Profiles, since the two are confounded in the design of the study. On the other hand, inter-rater and test-re-test reliability estimates would suggest that the effect of this confounding is not substantial (see Table 1, p. 16; Table 2, p. 17).

However, it is possible that direct comparisons between the VQSP and other studies of school effectiveness are inappropriate given the heavy reliance in the VQSP on teacher assessments of student achievement as opposed to students' scores on standardized tests or on public examinations. In essence, the decision to use the Victorian Profiles to obtain comprehensive, 'authentic' and curriculum-specific estimates of achievement involved certain trade-offs between reliability and validity, namely, the exchange of greater uncertainty with respect to reliability, for greater certainty with respect to validity. Just what is entailed by such trade-offs is not entirely clear, but as Moss (1994) argues, such issues go to the core of deep-seated beliefs about the ways in which we view research in education. Stringfield (1994b) observes that while there is a mature literature on reliability, validity is a less well-handled issue with discussion invariably sliding toward contemplations of the meaning of 'reality'.

Mention has already been made (see pp. 65-66) concerning the findings presented in Table 11 (p. 63), that the large proportion of variance in students' literacy progress at the class/teacher-level is residual variation that is inadequately accounted for by the fitted explanatory variables at that level, namely Literacy PD Program and Multigrade Class. Similarly (as also noted earlier), none of the measures of teachers' perceptions of their work environment were found to have significant direct effects on student progress. However, it is important to note that such findings are useful to the extent that they highlight a need for refocussing the prevailing school and educational effectiveness research agenda to one that is closer to students' experiences of schooling in class groupings, and re-examines class/teacher influences on progress in student achievement as advocated by Brophy and Good (Brophy, 1981, 1986; Brophy & Good, 1986; Good & Brophy, 1984), and more recently by Creemers (1992, 1994b), Slavin (1994, 1996), Schaffer et al. (1994) and Teddlie (1994).

From the explanatory models fitted, Attentiveness in the classroom had strong, positive effects on both elementary and secondary students' achievement progress. This finding is commensurate with a large clinical research literature showing strong relationships between students' learning outcomes and especially their attentive/inattentive behaviours (see references cited in Note 10). In particular, the evidence from studies investigating the impact of students' maladaptive behaviours at school highlight the importance of inattentiveness as a major factor having significant negative effects on students'
educational progress, especially in literacy. These studies reflect enduring concerns of school administrators, teachers, parents and health professionals of the extent to which the major characteristics of externalizing behaviour problems in the classroom, presently called disruptive behaviour disorders (i.e., attention deficit/overactivity and conduct disorders) – adversely affect students’ opportunities for learning and educational development. Students who typically exhibit inattentive, disruptive and maladaptive behaviours have been shown to be at risk of poor educational attainment, particularly in literacy. Moreover, in addition to the consequences for an individual, such behaviour problems in the classroom diminish educational opportunities for other students and contribute to teacher stress. The present results related to the importance of Attentiveness in the classroom in accounting for student achievement were illuminated by the finding that the negative effect of Critical Events on students’ progress is mediated essentially by Attentiveness (see Table 12, p. 65). Similarly, the finding that Prior Achievement had a strong positive effect on students’ Attentiveness, highlights the importance of early achievement for current and subsequent student behaviours in the classroom, as well as for their achievement progress.

Whereas there is considerable latitude within the VQSP for philosophical elucidation related to the conceptualisation and measurement of educational effectiveness, and the extent to which high validity can coexist with high reliability, on one issue there can be little doubt. Over and above class/teacher variation, the proportion of variance in students’ progress in Literacy and in Mathematics due to between-school differences is small (i.e., < 8%). In this regard, findings from the VQSP are more closely aligned with those from the foundation studies of school effectiveness by Coleman et al. (1966) and Jencks et al. (1972) rather than with the optimistic account of school effectiveness characterised by Edmonds (1979b) “five factor” model. As Scheerens (1992) and Reynolds et al., (1994) observe, there is a sense in which school effectiveness research appears to be re-visiting its origins, although with considerably more insight into the nature of those origins. Perhaps this is something to be expected. At least we can be encouraged by T. S. Eliot’s injunction: “We shall not cease from exploration and the end of all our exploring will be to arrive where we started and know the place for the first time” (1943, p. 59).

---


23 The link between academic underachievement and students’ externalizing behaviour problems has long been noted (see Sampson, 1966). For historical reviews of this interest and the related research, see: McGee, Share, Moffitt, Williams and Silva (1988); Hinshaw (1992a,b, 1994). A report specifically dealing with these issues has recently been published (Sept. 1996) by the British Psychological Society (BPS), entitled: Attention Deficit Hyperactive Disorder: A psychological Response to an Evolving Concept.

24 This finding applies to both boys and girls, but is particularly pronounced in the early literacy development of boys. For an overview of this research, see: Cantwell and Baker (1991); Davie, Butler and Goldstein (1972); Elkins and Izard (1992); Hinshaw (1992a,b, 1994); Keller et al. (1992); McGee and Share (1988); Maughan, Gray and Rutter (1985); Rowe (1991, 1995); Rowe and Rowe (1992a,b, 1993); Rutter (1985), Silver (1990).

25 See, for example: Brenner, Sörbom and Wallius (1985); Otto (1986); Wearing (1989).
References


Muthén, B. O. (1984), A general structural equation model with dichotomous, ordered categorical, and continuous latent variables, Psychometrika, 49, 115-132.


Rowe, K.J. (1991a). Students, parents, teachers and schools make a difference: A summary report of major findings from the 100 Schools Project - Literacy Programs Study. Melbourne, Vic: School Programs Division, Ministry of Education.

Centre for Applied Educational Research, The University of Melbourne


