NUI MAYNOOTH

Ollscoil na hÉireann Má Nuad

Issues in the Interpretation of PISA in Ireland:

A Study of the Content and Design of PISA with Comparative Analyses of the Junior Certificate Examinations and TIMSS (1 Volume)

Judith Cosgrove [BA, MA (Psychology)]

A dissertation submitted to the Department of Education at the National University of Ireland, Maynooth in fulfilment of the requirements for the degree of Ph.D.

Supervisors: Professor John Coolahan and Dr Thomas Kellaghan

October 2005

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I am indebted to Dr John Coolahan and Dr Thomas Kellaghan for their advice, guidance and encouragement in the course of working on this thesis. Thanks to school staff and students for participating in PISA; without their co-operation the survey could not have happened. Thanks to my parents, brother and sisters for their consistent interest in and encouragement with this project. Thanks to David Millar for much-needed support, and assistance with producing graphs in SPSS and Excel. Thanks to Gerry Shiel for general support, interest and encouragement. Thanks to Nick Sofroniou for advice on statistical treatments for clustered sample designs. Thanks to Christian Monseur for discussions on non-response bias. Thanks to the PISA Consortium, and to the OECD Secretariat, for their efficiency and support with technical and informational queries about PISA. Thanks to Seán Close and Elizabeth Oldham for their guidance and support, and for interesting discussions and collaborations on the mathematics curriculum in Ireland. Thanks to Tom Mullins for his support and advice on the English curriculum in Ireland. Thanks to Peter Archer for advice on and discussions about the social context effect and educational disadvantage and to both Peter Archer and Eemer Eivers for comments on an advanced draft of the concluding chapter. Thanks to the six individuals who completed the test-curriculum rating project. Thanks to the PISA National Project Managers for providing information on national curricular analyses and discussions of these. Finally, heartfelt thanks to Sheila Normanly, without whom this thesis would never have seen the light of day (with 100% certainty!).

SUMMARY

The results of the Programme for International Student Assessment (PISA) have been the subject of attention in media reports and ministerial speeches in Ireland since 2001. Yet, to date, there has been no examination of the appropriateness of the data it yields to inform educational policy. Analyses described in this dissertation were carried out to identify aspects of the design and interpretation of the PISA assessments of reading and mathematics which may be problematic and/or at odds with conclusions drawn. Three major questions are addressed. First, what does PISA tell us about achievements of students in Ireland? While PISA is intended to be used for educational improvement, the fact that it does not purport to assess school-based knowledge and skills could be problematic. In this context, the extent to which the PISA achievement measures are similar to, or differ from, the national curriculum (Junior Certificate syllabus and examinations) is examined. The PISA reading measure was found, by and large, to be compatible with Junior Certificate English. However, the diverging philosophies underlying PISA and Junior Certificate mathematics result in notable disparities between the two assessments which pose challenges in interpreting the results (although these disparities may serve a potentially valuable 'enlightenment' function and act as a trigger for curriculum review). Both PISA and the Junior Certificate Examination were found to be of very limited utility in describing performance at the low end of the achievement distribution. Analyses of non-response in the Irish datasets for PISA 2000 and 2003 reveal significant bias arising from student (but not school) non-response. This finding renders claims made by the OECD and Irish media about the characteristics of low achievers problematic and strengthens the argument that PISA is not well suited to describing characteristics of low achievers in Ireland. The second question considered is what PISA tells us about the equity of achievement outcomes in Ireland. Analyses of the between-school variance statistic were used to address this question since is widely cited as an indicator of educational equity. However, when inferences are being made on the basis of between-school variance, no cognizance is taken of how sample design and other characteristics of PISA might have impacted on its magnitude. This issue is of particular relevance in Ireland where studies have revealed large achievement differences associated with class allocation within schools. Furthermore, the nature of the achievement measure used (in particular, its curriculum sensitivity) would also appear relevant to the interpretation of the significance of between-school variance. Comparative analyses of the TIMSS 1995 and PISA data

indicated that between-'school' variance in Ireland is much larger when the sample design involves the selection of intact classes (in TIMSS) rather than on the basis of students' age (in PISA). Furthermore, school-dependent and curriculum-sensitive measures tend to be associated with higher between-school variance. The third question relates to what PISA tells us about the determinants of the achievements of Irish students. This question is of some significance since the impact of schools' social intake on student achievement, and the extent to which school practice variables explain achievement, are given prominence in the results of PISA and in other surveys both nationally and internationally. A comparison of multilevel models of Irish student achievement on PISA 2000, PISA 2003 and TIMSS 1995 using achievement on the international tests and the Junior Certificate Examinations suggests that the impact of social intake is larger when surveys use a sample design based on intact classes, while school-dependent and curriculum-sensitive measures may be more sensitive to school practice variables. In the conclusions, some policy implications are described, the limitations which these findings place on the interpretation of results considered, some improvements to the design of PISA that may overcome some of the limitations suggested, and areas for future research proposed.

LIST OF ABBREVIATIONS AND ACRONYMS USED

ACER	Australian Council for Educational Research
EJCPS	Junior Certificate Performance Scale for English
ESCS	Economic, Social and Cultural Status
ETS	Educational Testing Service (USA)
FIMS	First International Mathematics Study
IEA	International Association for the Evaluation of Educational Achievement
INES	Indicators of Education Systems
IRT	Item Response Theory
JCE	Junior Certificate Examination
JCPS	Junior Certificate Performance Scale
MJCPS	Junior Certificate Performance Scale for Mathematics
NAE	National Academy of Education (USA)
NAEP	National Assessment of Educational Progress (USA)
NIER	National Institute for Educational Research (Japan)
OECD	Organisation for Economic Co-operation and Development
OTL	Opportunity To Learn
PGB	PISA Governing Board
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
SES	Socioeconomic Status
SIMS	Second International Mathematics Study
TIMSS	Third International Mathematics and Science Study

CHAPTER 1. BACKGROUND TO INTERNATIONAL ASSESSMENTS OF EDUCATION, AND DESCRIPTION AND EVALUATION OF PISA'S OBJECTIVES AND DESIGN

1.1. Introduction

The results of the Programme for International Student Assessment (PISA) have been in the public domain since December 2001. Yet, to date, there has been no formal, academic study of PISA in Ireland. The background from which it emerged may be traced back to the 1950s and considerable developments have taken place since the earlier comparative studies of educational achievements. This chapter aims to fulfil the purpose of providing an historical, interpretative overview of PISA. More specifically, it provides a context in which to consider PISA; describes its design and objectives; offers a broad evaluation of the survey; demonstrates its place within the Irish system for monitoring education; and considers claims made by the Irish media and politicians about the results. First, origins of international assessments of educational achievement are reviewed. Second, concerns with educational standards are traced from the 1950s to the inception of PISA in 1997. Third, early international studies are reviewed. It is shown how these have shaped recent survey designs conceptually and thematically; more recent developments have related largely to methodological advances. Fourth, the structure and functions of the OECD are described and the role of PISA within the broader activities of the OECD are reviewed. Fifth, PISA's aims, assessment frameworks, and survey design are described. Sixth, some issues in interpreting the results of international assessments are noted and PISA is evaluated with respect to these. Finally, the role of PISA in the national system for monitoring educational outcomes and its importance with respect to national educational policy are described.

1.2. Educational Outcomes

This section provides a description of the main concepts and issues pertaining to educational outcomes, and traces the growth in concern with educational standards from the 1950s.

1.2.1. Concepts and Main Issues Associated with Educational Outcomes

The assessment of educational outcomes can be considered within the broader framework of outcome evaluation. A number of features of outcome evaluation have been identified (Kellaghan & Madaus, 2000). First, it usually refers to activities that are designed to measure the effects or results of programmes (such as students' achievement) rather than their inputs or processes. Second, judgements with respect to a standard are commonly made, and the idea of 'excellence' is evident. For example, categorical proficiency levels, which use characteristics of the test items as a basis for describing the likely skills of students scoring at various points on the achievement scale, are commonly interpreted with respect to standards in assessments of educational outcomes, even though the standards associated with the levels are not explicit (e.g., Kellaghan, 2001). Thus, it can be left to policy makers and the public to interpret and evaluate these results. There are obvious dangers to this for both epistemological and technical reasons, and the scientific methods used to set standards may be "sublimated to serve unscientific ends" (Cizek, 2001, p. 14). Third, most usually, but not necessarily, the focus is on outcomes on completion of a programme. In the case of PISA, the assessment is designed to examine the achievements of students who are at or near the end of compulsory schooling (OECD, 2001b; 2004c). Fourth, it is not usual to describe what is actually happening in a programme (since the design of the assessment usually takes the form of a cross-sectional survey), although the kind of information obtained will generally be chosen to reflect programme activities. Fifth, efforts are often made to relate outcomes to contextual factors or antecedents, and such data are commonly collected through ancillary questionnaires. Sixth, outcome evaluation may be once-off or may involve monitoring over time. In recent international assessments, the practice has been to monitor educational outcomes over time whereby surveys are repeated every three years (in the case of PISA), every four years (the Third International Mathematics and Science Study, TIMSS), or every five years (Progress in International Reading Literacy Survey, PIRLS).

National and international assessments are the two main ways in which information is obtained about educational outcomes in the context of outcome evaluation. Both types of assessment are a feature of many education systems, including Ireland (Section 1.8 describes Ireland's system for monitoring educational outcomes), and these commonly assess students' first language and mathematics at primary level. Science is less

commonly assessed (Kellaghan & Madaus, 2000). International and national assessments follow similar procedures (e.g., sample design and selection, data collection, scaling of achievement data, explanatory analyses of achievement) but differ from each other in several obvious ways (e.g., with respect to instrument construction and the possibility of cross-national comparative analyses). The main advantage of international assessments is they allow an indication of where students in a country stand relative to students in other countries (Greaney and Kellaghan, 1996; Kellaghan and Madaus, 2000).

Key to the interpretation of national educational outcomes is their measurement (assessment), monitoring (systematic and regular procedures for the collection of data about important aspects of education at national, regional or local levels), and evaluation (collection and interpretation of evidence, leading to a value judgement) (Greaney & Kellaghan, 1996).

An education outcome may be considered as a type of educational indicator, which is a numeric value used to describe policy-relevant statistics that contain information about the status, quality or performance of an education system. Educational indicators should have the following characteristics: they are quantifiable; can be judged against a standard or criterion, be it norm-referenced, self-referenced, or criterion-referenced; are generally viewed as important; describe conditions that are generally amenable to improvement; are collected frequently enough to allow monitoring; allow comparisons of subgroups; and are based on a theoretical model of how the education system is thought to work (Greaney & Kellaghan, 1996).

In a country in which public examinations occupy such an important role, it may be worthwhile to draw attention to the differences between them and international/national assessments. In a nutshell, public examinations are used to determine whether *individual* students possess certain knowledge and skills, while national/international assessments are used to evaluate the education system *in general*, or a clearly-defined part of it. There are consequently differences in scoring and reporting. In public examinations, the main result of interest is how students perform with reference to one another, with particular attention to higher achievers. In the case of national/international assessments, on the other hand, the main aim is to be able to say

something about the knowledge and skills of groups of students, and results are often reported in terms of performance criteria (e.g., proficiency levels) on a standardised scale which aims to discriminate over a broader range of ability. Public examinations and large-scale national/international assessments also tend to differ according to whether they are high stakes (where the results have important and direct consequences for those who take the test) or low stakes (where results are of relatively little importance or have only indirect consequences for those who take the test) are attached to performance. These differences may impact on student effort and motivation and preparedness for the test (Klein & Hamilton, 1999). Greaney and Kellaghan (1996) argue against the use of public examinations data to monitor education systems, due to the fact that they are used as a basis for allocating individuals rather than providing information about the system as a whole; are not standardised; and are prone to fluctuations relating to test content, marking and changes in the population attempting the examination.

1.2.2. Early Concerns with Measuring Educational Outcomes in an International Context

According to Husén and Postlethwaite (1996), concern with the measurement of educational outcomes was being expressed almost at the outset of the establishment of UNESCO in 1945 (see UNESCO, 2003). A number of meetings were held in the mid-1950s at the UNESCO Institute of Education (UIE), Hamburg, at which educational psychologists, mostly psychometricians, discussed problems of school and student evaluation. This culminated in 1958, in a meeting once again in Hamburg of a small group of educational psychologists and sociologists, most of them from the US and UK, to consider undertaking a study of "measured outcomes and their determinants within and between systems of education" (Husén & Postlethwaite, 1996, p. 129). The meeting might be considered the point at which the International Association for the Evaluation of Educational Achievements (IEA) was founded. It was agreed that the practice at that time – using graduation rates as a proxy for the productivity of a system – was too simplistic, and lacking a qualitative dimension, since it did not capture what students actually learned. It was agreed that there was a need to focus on educational outcomes (internationally valid measures of achievement) which would provide countries with a comparative framework, later to become a major issue in considering international

competitiveness in knowledge-based economies in an increasingly globalised context (Kellaghan & Greaney, 2001).

A comparative approach, it was hoped, could capitalise on the variability across education systems in the sense that systems could be likened to "one big educational laboratory" (Husén, 1973, p. 22). The research perspective had a structural functionalist approach or philosophy, characterised by an emphasis on empirical, quantitative analytic methods (Gibbons & Sanderson, 2002; Husén, 1997; Welch, 1999). Outcomes were selected if considered to be of policy relevance, and presumably, if they were measurable within the structural functionalist framework.

Governments at that time were also concerned that labour force needs, particularly in the area of technology, would not be met, and it was thought that one way to help meet this need was through education. Kellaghan and Madaus (2000) identify the publication of the Coleman report (Coleman et al., 1966) as a pivotal point at which public attention moved from resources or inputs to a focus on outcomes as measured on tests. They note that there was a growth during the 1960s in corporatist approaches to government administration, and many ideas borrowed from the business world (e.g., strategic and operational planning, use of performance indicators, focus on deliverables and results, accountability systems based on results). The accounting community was also exerting an increasing influence on government administration, reflected in performance audits, value for money audits, etc. They also identify a shift in the use of assessments as a localised tool to assist decision-making in instruction to a policy tool used in centralised, high-stakes policy-making and accountability monitoring. These changes, coupled with a growth in demand for public services and the funding of social programmes, resulted in a greater emphasis on efficiency of programmes and more selectivity in the support of various programmes. Hence the notion that international studies could provide information on "optimal conditions for human development that could be used as a basis for educational policy" (Kellaghan, 1996, pp. 143-144) must have been quite appealing. As Foshay et al. (1962, p. 7) put it: "If custom and law define what is educationally allowable within a nation, then educational systems beyond one's national boundaries suggest what is educationally possible."

Postlethwaite (1999) has identified four functions of comparative educational research, all of which are potentially relevant to the development of government policies on education. First, the identification of what is happening in other countries might be used to guide improvements to the national education system. Second, an analysis of similarities and differences in goals, structures and achievements (i.e., differences in inputs and processes and how these relate to outcomes) is of potential practical significance. Third, given the level of speculation in educational research about 'what affects what' (Kaiser et al., 2002, p. 632), the quantification of the extent to which, for example, school and teacher variables are associated with achievement, is of interest. A fourth is the identification of general principles concerning the effects of education.

Kellaghan (1996) has also discussed some potential uses of these studies. First, they serve to identify aspects of a system (such as curricular content or achievement outcomes) that are at odds with other systems. Second, the findings can contribute to the identification of optimal conditions for development, relevant to many decisions about the deployment and nature of resources. Third, results might have a slower, but nonetheless just as important, philosophical or epistemological impact in that they can reveal assumptions about what schools try to achieve, through an analysis of what they actually achieve and a discussion about what it is possible to achieve. This has been termed the 'enlightenment function' (see also Postlethwaite, 1999). Fourth, the studies serve as mechanisms of accountability (but how these mechanisms operate, and which bodies will be held accountable, is often far from clear). Results of the studies can also be used to monitor standards (providing 'objective' evidence about claims of groups such as employers that educational standards are rising or falling); to introduce realism into debates about appropriate and attainable achievement levels; to direct efforts at raising achievements; and to increase public awareness and inform public debate.

It was, however, two decades before interest in the findings of such studies gained significant momentum internationally. The International Indicators of Education Systems (INES) project of the OECD was not established until 1988 (Kellaghan, 1996). This is surprising, given the relative success of the first pilot study and first international assessment carried out by the IEA (Foshay et al., 1961; Husén, 1967a, b). The early studies, however, were generally under-funded, and results were not reported in a timely manner (Husén & Postlethwaite, 1996). This can in large part be attributed to a lack of

government interest and support. This seems ironic given that, in the report on the First International Mathematics Study (FIMS; reviewed in the next section), Husén (1967a) emphasised that the intent of the studies was to address the perceived need for policy-making and educational planning. Furthermore, the work of the IEA was innovative in three respects in providing data which were previously unavailable. First, data on outputs as well as inputs and processes were gathered. This represented a new kind of information since it was in a comparative context. Secondly, the studies capitalised on variability across educational systems, allowing a broader range of educational conditions to be studied. Thirdly, the use of an empirical, quantitative approach was new in the field of comparative studies of education (Kellaghan, 1996).

The lack of support can be traced to the fact that, during the 1970s, there was a decrease in the belief among OECD member countries in the value of investment in education, as well as a trend to move away from macro-level educational planning, as the standards-based reform movement in the US collapsed (Husén & Postlethwaite, 1994; Kellaghan, 1996). This lack of support can also be interpreted within a broader *malaise* with science and technology which were not showing the way to overcome barriers to the attainment of peace, wealth and happiness (Gibbons & Sanderson, 2002). Welch (1999, p. 35) has commented that the oil crisis of the late 1970s resulted in recessions and high rates of unemployment. This, coupled with the deregulation of many economies resulted in a "... general decline in government activity and intervention in social and economic affairs". A second reason for the decline in support, suggested by Postlethwaite (1999), is that at the outset a large amount of work of the IEA was done by a relatively small number of people which meant that personnel resources as well as financial ones were scarce. Further, although the work of the IEA began in 1958, it was not made an official body until 1966 (Husén & Postlethwaite, 1996). Thus, the systematic funding of such studies was not built into the structure of the IEA at the outset, and the studies were not backed at government level, as is the case with OECD surveys.

1.3. Early Studies

This section reviews the first international assessment of educational achievement, and its preceding pilot survey, in order to identify some of the aspects of the survey design that inform the designs of current assessments, to highlight some conceptual and methodological concerns that are still relevant in the interpretation of contemporary

survey results, and to identify some developments (mainly methodological) since the early surveys.

1.3.1. The First International Pilot Survey, and the First Full-Scale International Survey

A pilot study of the measurement of educational outcomes was carried out by the IEA in 1961. The aims of the pilot were to see whether indications of "intellectual functioning" (Foshay et al., 1962, p. 7) could be deduced from responses to short-answer tests, and to examine the feasibility of conducting a large-scale international study of educational outcomes. The assessment covered four subject areas (reading comprehension, mathematics, science, and geography), as well as non-verbal aptitude (abstract logical and analogue reasoning). Most items were multiple-choice, though the tests included a few constructed-response items. The target population was age- rather than grade-based (pupils aged 13 years 0 months to 13 years 11 months on the first day of a country's school year). The particular age group chosen relates to the fact that, at the time, in many countries, pupils left school by the age of 14. The sample design was poor since convenience sampling was used and the numbers sampled were small (300-1700 students per country). Many countries selected a group of students from one region in a country deemed 'representative' of the population, which is problematic if one wishes to generalise the results from a region to the country. The samples therefore lacked precision and were not representative. Student exclusions were noted but not documented in detail. Testing conditions were not comparable since no time limit was placed on testing time. Questionnaires for students and their teachers were administered and 15 student and school background indicators developed.

It might be noted that the reliabilities of the mathematics and reading tests (both .81) were higher than that for geography (.70) and science (.62). Pairwise correlations also reveal that the mathematics and reading results are more comparable across countries than results in the other two subjects (.87 for both mathematics and reading versus .68 for geography and .72 for science). Postlethwaite (1999) has commented, however, that coming to an agreement of what constitutes a test of mathematics and a test of science is more complex than a test of reading. Reading is a more generic skill which cuts across curriculum areas and practiced in many contexts. The learning of mathematics and

science in contrast is largely confined to school contexts and hence more dependent on school curricula.

Results were reported in terms of averages and standard deviations, overall and by country. Between-country variance, relative achievement by subject, sex differences in achievement, and analyses of background characteristics such as type of school attended and socioeconomic status were also reported. These analyses arguably still form the core of current analyses of international assessments of educational achievements.

One chapter in the report of the first survey by Walker dealt with teachers' ratings of the items (other than the reading items) and the proportion of variance explained by the relative emphasis and exposure of students to the content of the tests (as indicated by teachers). The proportion of achievement variance accounted for by ability was much higher than that accounted for by curricular exposure or emphasis, however. This analysis formed the blueprint for measures of curricular exposure of students to the test items, later to be termed 'opportunity to learn' (OTL) (see Floden, 2002).

According to Husén and Postlethwaite (1996), many of the outcomes of the pilot were of both academic and practical value, and the main conclusion drawn was that it was feasible to conduct a large-scale international assessment. Foshay et al. (1962) commented:

> The data obtained, even under the restrictions that inevitably arise, can be analysed fruitfully. And, by extension, we think we have shown that it is possible to introduce a large empirical element into comparative education – an element only slightly present in the field until now. (p. 19)

A proposal to carry out a large-scale survey in mathematics (only called the First International Mathematics Study or FIMS upon administration of the second mathematics study, SIMS) was drafted by a number of individuals involved in the IEA pilot, and submitted by Bloom and Anderson to the US Office of Education (USOE) (Husén & Postlethwaite, 1996). The results of FIMS were reported in two volumes (Husén, 1967a, b).

According to Robitaille and Travers (1992), the choice of mathematics as the subject of enquiry for the first international study was influenced by convenience. Many of the 12 participating countries were concerned with improving scientific and technical education. In addition, the 'new mathematics' movement (see, for example, Oldham, 1980a, b; 1989; 2002) resulted in increased interest in comparisons of curricular changes across the participating countries.

The problems with the sample design of the pilot survey were largely overcome by the involvement of Peaker of the National Foundation for Educational Research (NFER), England, who developed a sampling manual, to which participating countries adhered (Husén & Postlethwaite, 1996). Another notable development, by Walker, was the creation of an OTL measure (a precursor of which was documented in the 1962 report). This was developed on the basis that, while the cognitive tests could be developed with reference to the curriculum documents and textbooks of participating countries, there might be a difference between what these contained/laid out and what teachers actually taught (later, this distinction was referred to as intended and implemented curriculum).

Two age groups were assessed: 13-year-olds (both age and grade samples; i.e., all pupils aged 13.0 to 13.11 years at the date of testing, and all pupils at the grade level at which the majority of 13-year olds were found; populations 1a and 1b, respectively) and students in their final year of schooling (those who were studying mathematics as an integral part of their schooling versus those who were studying mathematics merely as a complementary part of their schooling; populations 3a and 3b, respectively) (Husén, 1967a).

Although no explicit definition of mathematics is given in reports, the framework from which the assessment was developed seems to have been moderately comprehensive and based on consensus between countries as to what constitutes mathematical knowledge and skills. The construction of the test was informed by reports prepared by each national centre outlining the content and objectives of the national mathematics curricula and also including, in some cases, draft test items. An item was deemed fit for inclusion in the assessment if it was considered appropriate in at least three of the 12 participating countries.

Results were reported in terms of total test scores, lower mental process, higher mental process, verbal mathematics, and computational mathematics. Items were also analysed by traditional topic areas (arithmetic, algebra, geometry, calculus, sets and logic). The parallels between the processes identified in FIMS and the three competency classes described in the PISA framework (described in Section 1.6) are notable. Both distinguish between routine reproduction of mathematical procedures and more creative problem-solving in novel contexts.

1.3.2. Issues Identified in Early Surveys, Common Themes, and Methodological Advances

The IEA pilot and FIMS established many aspects of the designs of more recent international surveys, such as the Third International Mathematics and Science Study (TIMSS) and its two follow-up surveys, TIMSS 1999 and TIMSS 2003 (sometimes referred to as TIMSS-repeat and TIMSS-trends, respectively) (e.g., Beaton et al., 1996a, b; Martin et al., 2000a; 2004; Mullis et al., 2000; 2004), the Progress in International Reading Literacy Survey (PIRLS; Mullis et al., 2003a), and PISA (e.g., OECD, 2001b; 2004c, d).

First, all recent surveys are replete with sampling manuals and other procedural documentation to help to ensure the representativeness and precision of the samples, and cross-national comparability of operational procedures such as test administration, timing of test sessions, and adaptations to test and questionnaire items.[1] In all surveys, the sampling standards are clearly specified and each country's sample is evaluated with respect to these standards, and flagged in international reports if it falls short of the standards, or in some cases, omitted from the reports. However, the debate with respect to whether an age-based or grade-based design is more appropriate to the aims of the survey is ongoing (Postlethwaite, 1999; Smithers, 2004) and consists of much the same issues as identified in FIMS. In the report for FIMS, it is noted that the problem of defining the populations to be tested was "extremely complicated and took up a great deal of time. ...Great difficulties were experienced in selecting populations which were, in fact, comparable in terms of their place in the educational structure" (Husén, 1967a,

[1] For examples, procedural documentation for PISA 2000 can be accessed at http://www.pisa.oecd.org/findDocument/0,2350,en_32252351_32236159_1_119669_1_1_1,00.html; for PISA 2003, at http://www.pisa.oecd.org/ findDocument/0,2350,en_32252351_32236173_1_119669_1_1_1,00.html.

p. 45). The 'grade versus age' debate is taken up further in Section 1.7, where consequences for the interpretation of the results are considered in more depth.

Second, each of these has an assessment framework which describes in detail the skills and concepts associated with the domain(s) assessed and how these are mapped onto the test specification (e.g., Campbell et al., 2001 [PIRLS]; Martin, Gregory & Stemler, 2000 [TIMSS 1999]; Mullis et al., 2003b [TIMSS 2003]; OECD, 1999c, 2000a, 2003b [PISA 2000 and 2003]; Robitaille et al., 1993 [TIMSS 1995]). A measure of OTL has been included in all assessments of mathematics and science with the exception of PISA. Along with the comparability of the target populations and the sample design chosen, the content and design of the test instruments has been identified as a particular area of difficulty in international comparative studies of educational achievement in terms of its relevance to national policy priorities and fairness to students, depending on its similarity with/difference to what they are usually taught/assessed (Beaton et al., 1999; Postlethwaite, 1999). In the report for FIMS, the difficulties involved in producing an internationally equivalent instrument (in terms of match to the curriculum across countries) were noted:

> ... the task of preparing a battery of tests that could be used in common by the various countries and that at the same time would have high curricular validity for all countries and programs was regarded as impossible. (Husén, 1967a, p. 84.)

Third, the themes considered in explanatory analyses of achievement remain broadly similar, entailing a consideration of student- class- and/or school-level variables, and including a consideration of social background. However, multilevel modelling techniques are now commonly used, and preferred to the regression techniques associated with the earlier studies (e.g., Martin et al., 2000b; OECD, 2005c). A broad review of multilevel models as applied to educational achievements is included in Section 1.7.

Fourth, comparisons of student achievement by sub-areas of the domain in question are common to all international assessments, but the manner in which student achievement is aggregated has been transformed dramatically from a simple averaging of correct responses expressed as percent correct to a scaled achievement score created within an Item Response Theory (IRT) framework (Adams, Wilson & Wang, 1997). In recent

surveys, the multidimensional random coefficients multinomial logit model, which allows partial credit items to be included as well as items with right-wrong responses, is preferred. However, there is some debate as to whether a Rasch model (one-parameter model, whereby item difficulty is presumed independent of test-takers' ability) is preferable to a three-parameter model (whereby item difficulty is dependent on ability) in the scaling of international achievement data (see Hambleton et al., 2005). Regardless of whether the underlying model has one or three parameters, however, it has the advantage of placing item difficulty and student ability on the same scale and the score a student receives is independent of the particular set of items attempted. Consequently, it is possible to include a broader range of items than could be attempted by a single student through the use of a rotated booklet design, whereby each student attempts only a subset of items. Common blocks of items are used to establish psychometric links across all of the test forms. A second advantage of IRT is that it allows a description of skills in terms of characteristics of test items that students are likely to be able to achieve a correct response on. This property of the achievement scales has been exploited with the construction of categorical proficiency scales, whereby cutpoints are applied to the achievement scale and, through a consideration of the items at each level on the scale, the likely skills of students at each level on the proficiency scale. This is an advantage since it allows a qualitative interpretation of a test score and it can also be used as shorthand to describe the distribution of achievement and set educational standards. However, as I show in Section 1.7, the interpretation of proficiency levels is a controversial aspect of recent surveys.

1.4. Background to the OECD[2]

Ireland is one of the 20 original member states of the OECD, which was established in December 1960 with the signing of the Convention on the Organisation for Economic Co-operation and Development. Since then a further 10 countries/regions have joined and the OECD also has links with a further 70 countries/regions with 'transition and emerging market economies'. Its origins are in the Organisation for European Economic Co-operation (OEEC), which was established in April 1948 at the Conference for European Co-operation, as a result of the work of the then US Secretary of State,

[2] All material in Sections 1.4 is based on information provided on the OECD website, at http://www.oecd.org.

George Marshall. The OEEC sought to establish a permanent organisation to maintain work on a joint post-war recovery programme.

The OECD Convention (1960) is based on the belief that economic strength is the key to the attainment of the purposes of the UN, preservation of individual freedom, and enhancement of a nation's general well-being. Further, the Convention holds that the best way to achieve this is through co-operation, recognising the increasing economic interdependence among member countries. The motto of the OECD is *building partnerships for progress.*

The aims of the OECD are to promote policies designed to:

- achieve the highest sustainable economic growth and employment and a rising standard of living in member countries, while maintaining financial stability, and thus contributing to the development of the world economy;
- contribute to sound economic expansion in member as well as non-member countries in the process of economic development; and
- contribute to the expansion of world trade on a multilateral, non-discriminatory basis in accordance with international obligations.

There is a high emphasis in the work of the OECD on the development and production of valid and reliable economic, educational, and social statistical indicators and the OECD has produced an enormous body of statistical reports and economic, educational, and social reviews.

The role of education in a nation's economic well-being, according to the OECD, is paramount. Skills need to be transferable and continually updated, and are a form of human capital – probably the key form of human capital. The OECD (1998) defines human capital as "the knowledge, skills, competencies and other attributes that are embodied in individuals that are relevant to personal, social and economic well-being" (p. 9). Thus the availability of good measures of outputs associated with education systems is important. Kellaghan and Greaney (2001, p. 95) note that "assessment is seen as having a major role to play in ensuring that the outcomes of education and training are those that the economy needs". The OECD argues that the current economies are knowledge-based, perhaps knowledge-driven. It further states:

Both individuals and countries benefit from education. For individuals, the potential benefits lie in general quality of life and in the economic returns of sustained, satisfying employment. For countries, the potential benefits lie in economic growth and the development of shared values that underpin social cohesion. (http://www.oecd.org)

1.5. The Collection of Educational Indicators by the OECD

1.5.1. The Establishment of INES

It was noted in Section 1.3 that there was a decline in interest in international comparative education and that INES was not established until 1988 despite the success of the IEA pilot and FIMS (and the innovative nature and policy relevance of the research). However, in the 1980s, a change in the official positions of OECD member countries was evident. This has been attributed to higher school completion rates in many countries. As already noted, reliance on graduation rates as an indicator was no longer viewed as sufficient, and some degree of quality assurance in the output was also needed. Costs per student as well as overall costs were rising, so concerns and questions were being asked about cost-effectiveness and accountability (Husén & Postlethwaite, 1996; Kellaghan, 1996). The political agenda of the USA relating to economic competitiveness was also instrumental in the establishment of INES and the establishment of PISA was a logical progression from the work of INES.

The publication of the first *Education at a Glance* (OECD, 1992b), the OECD's annual publication of educational indicators (which, along with the companion volume, *Education Policy Analysis,* form the principal outputs of the INES project), was aided by a significant financial contribution from the US Department of Education through the National Center for Educational Statistics (NCES) (this is noted in the preface to the 1992 *Education at a Glance*). A description of the events leading to the establishment of INES (OECD, 1992a) states that the US Department of Education initiated an international conference on education indicators which took place in 1987 in Washington DC. At that conference, the OECD was invited to undertake developmental work on a comparative set of education indicators. This was confirmed at the International Conference on the Evaluation of Education Systems in Poitiers, France, in 1988. As described in OECD (1992a):

Both conferences confirmed the urgent need for better and more comprehensive information about the outcomes of education. The current political debate is characterised by growing concern with qualitative aspects of education systems, over and above the traditional management questions that arose as a result of the massive expansion of education systems in the post-war period. ... It [also] emanates from new attitudes that have influenced and changed educational expectations. At the same time, evaluation mechanisms more sensitive to qualitative aspects of public service institutions are being developed ... And is forcing education authorities to rethink the issues...(p. 8)

According to Kellaghan's (1996) discussion of the work of the IEA, the international comparisons themselves were also creating momentum. A rather dramatic report of the US National Commission on Excellence in Education entitled *A Nation at Risk* opens by stating:

Our nation is at risk. Our once unchallenged pre-eminence in commerce, industry, science, and technological innovation is being overtaken by competitors throughout the world. ... What was unimaginable a generation ago has begun to occur – others are matching and surpassing our educational attainments. If an unfriendly foreign power had attempted to impose on America the mediocre educational performance that exists today, we might well have viewed it as an act of war. (1983, p. 5)

The link between educational standards, freedom and competitiveness in a global market economy is clear here: "Knowledge, learning, information, and skilled intelligence are the new raw materials of international commerce" (National Commission on Excellence in Education, 1983, p. 7). Data from IEA studies were used in *A Nation at Risk* to demonstrate the decline in academic performance of American students since the 'Sputnik Era'. For example, it is stated that in comparisons of performance on 19 academic tests over recent years, the US came last relative to other industrialised nations on seven occasions.

Another example of the gathering momentum of these comparisons in the US is discussed by Spaulding (1989), who commented that comparative study of education was an agenda for chiefs of state by the middle of the 1980s. He cites the example of a meeting between Presidents Reagan (US) and Nakasone (Japan) in 1983 and as a result, the setting up of a task force by each state to study the education system of the other, which was published by 1987. He comments: "IEA figures showing that Japanese

achievement scores in mathematics are higher than in the US undoubtedly motivated, in part, presidential interest in the studies" (p. 9).

According to the National Academy of Education (NAE), the commitment of the US to the development of a comprehensive programme of international comparisons received approval at the highest level of government when it was announced, at the 1989 Education Summit by President Bush and the nation's governers, that six broad education goals were to be achieved by the year 2000 (see Kellaghan, 1996). One of the six goals was that Americans were to be the first in the world in mathematics and science by the year 2000. The development of these standards was endorsed by Congress early in 1992 (National Council on Education Standards and Testing, 1992).

Thus, INES was established in an alarmist climate and concerns with educational outcomes within the world's major market economy at a time when competitiveness within a globalised community was gaining importance. The next section considers the structure of INES and locates PISA within it.

1.5.2. The Structure and Work of INES[3]

The INES project is concerned with indicators for cross-national comparisons of education systems. It develops, collects, analyses, and interprets indicators for international comparisons disseminated through its annual publications, *Education at a Glance* and *Education Policy Analysis*. It also provides a forum for international co-operation and exchange of information about methods and practices which aims to facilitate development in assessment methodology and practice, and enhance understanding of the use of indicators in policymaking. The work of INES is carried out by three Networks (A, B, C) and a Technical Group, each of which is focused on a different charge and which is chaired by a particular member country based on the relative financial contributions of member countries (with the highest-contributing country chairing the network in question).[4] Each of these is described in brief below since their work is interlinked, although the most relevant to a consideration of PISA is Network A.

---

[3] Information in Section 1.5.2 is based on information on the NCES website at http://nces.ed.gov/surveys/international/ INES/

[4] Shiel, personal communication, August 22, 2005.

The mission of Network A, chaired by the USA, is to develop indicators of learning outcomes, relating both to achievement and to social, emotional and attitudinal outcomes. PISA forms a core activity of Network A. Network B, chaired by Sweden, develops indicators of socioeconomic outcomes of education, such as education and work status of youth, labour force participation/unemployment by education level, and education and earnings by employment category. Network C, chaired by the Netherlands, develops indicators on the learning environment and organisation of schooling [e.g., intended curriculum time by subject area, teacher's working/instruction time, teacher salaries, student admission, placement and grouping policies, and decision-making in education systems which appear in *Education at a Glance 2004* (OECD, 2004b)]. Network C has carried out two surveys in the past few years, the *International Survey of Upper Secondary Schools* (ISUSS) in 2002 (OECD, 2004a), and the *Locus of Decision-Making* in 2003. It is planned to publish indicators based on this survey in *Education at a Glance 2005*. Network C is also exploring ways to improve the current system-level indicators on teachers and in developing a teacher workforce survey with an optional link to teachers in schools participating in future cycles of PISA.

The INES Technical Group is responsible for providing the majority of statistical data used for indicators of participation, access, human and financial resources, and school completion in *Education at a Glance*. This includes both developing a conceptual framework for reporting on education systems and conducting methodological studies to confirm the validity and comparability of these data. Indicators produced by the Technical Group include size of the school population, educational expenditure per student, support for students through public subsidies, and access to and participation in tertiary education.

1.5.3. The Establishment of PISA

At the time of the publication of the first *Education at a Glance* (OECD, 1992b), there was a mismatch between the definitions and available indicators, since most of the available indicators focused on inputs (OECD, 1992a). Bottani and Tuijnman (1994) note that (in the early 1990s) data on outcomes were difficult to obtain. They distinguish between outcomes of students, systems and labour markets, and comment that "All three areas are problematic, but the most difficult is by far student performance" (p. 68).

Around that time, the OECD relied on data collected by other agencies for other purposes as indicators of student achievement (namely, the International Assessment of Educational Progress [IAEP] and TIMSS). This did not match the OECD's needs since the available data were sporadic, infrequent, not available for all countries, and available in only a limited number of subject areas.

A decision was made by the OECD in September 1997 to establish its own procedures for data collection to increase comparability and scope of the indicators. The resulting programme was to produce student achievement indicators on a regular basis. This gave rise to the Programme for International Student Assessment, or PISA.[5] The PISA Governing Board (PGB) (originally the Board of Participating Countries; BPC) was set up in September 1997 when it was decided to implement a decentralised programme for producing student achievement indicators on a regular basis. The function of the PGB is to supervise the implementation of this Programme. The PISA results were reported in *Education at a Glance* for the first time in 2002 (OECD, 2002a).

1.6. Overview of PISA

1.6.1. Aims

1.6.1.1. Description

PISA's aim is to measure how well young adults approaching the end of schooling are "prepared to meet the challenges of today's knowledge societies" (OECD, 2004c, p. 20). Knowledge and skills that are deemed important for the present and future lives of 15-year-olds as individuals and as members of society have been identified by international panels of subject domain specialists. As noted in Section 1.5, PISA stems from the INES project. Indicators gathered through PISA include not only achievement outcomes but also background variables which are thought to be related to achievement. These are intended to complement other indicators gathered by INES, including the financial and human resources invested in education, access to education, and the learning environment in schools. PISA should be viewed in the context of current interest of governments in human capital. To the extent that the PISA assessment domains measure the knowledge and skills required for future adult life, performance on these domains may be interpreted as indicators of human capital.

[5] See http://webnet3.oecd.org/OECDgroups/

PISA produces:

1. A basic profile of student knowledge and skills among students at the end of compulsory schooling. Specifically, the OECD reports country average achievements with reference to the OECD average; the distributions of achievement in terms of percentile points and the percentages of students at various points on categorical proficiency scales; measures of the dispersion of achievement between students (the standard deviation) and schools (the percentage of achievement variance that is between schools); and illustrates the results with a number of sample tasks and student performance on these tasks.
2. Contextual indicators relating results to student and school characteristics; i.e., based on student-level and school-level data collected via student and school questionnaires, which range from student social background to school funding sources. The precise indicators collected/derived vary from cycle to cycle depending on the major domain and on policy priorities agreed on by the PGB, but a core focus of the analyses is social equity: the extent to which social background indicators at the student and school levels are associated with achievement.
3. Trend indicators showing how results change over time. Trend indicators were available for the first time in PISA 2003, where within-country comparisons of 2000 and 2003 results of both average performance and the scores of students at various percentile points were reported.
4. A knowledge base for policy analysis and research. The OECD places the PISA database, comprising responses to individual test and questionnaire items, as well as scaled achievement scores, composite variables, and supporting technical documentation on its website at http://http:pisa.oecd.org (see also OECD, 2003b, p. 8).

According to the OECD (2004c, p. 22), PISA can be used by countries to:

- gauge the literacy skills of their students in comparison with students of other participating countries;
- establish benchmarks for educational improvement, in terms of the performance of other countries, or their capacity to provide high levels of equity in educational outcomes and opportunities; and

- understand relative strengths and weaknesses of educational systems.

1.6.1.2. Justification

The approach underlying PISA is quite different to earlier surveys, particularly those of the IEA (OECD, 1999c; Smithers, 2004). The term 'literacy' is tagged to each assessment domain, a concept more usually reserved for the domain of reading. Further, in contrast to previous surveys, the OECD states:

> Although the domains of reading literacy, mathematical literacy and scientific literacy correspond to school subjects, the OECD assessments will not primarily examine how well students have mastered the specific curriculum content. Rather, they aim at assessing the extent to which young people have acquired the wider knowledge and skills in these domains that they will need in adult life. (OECD, 1999c, p. 9)

Describing the PISA approach as 'broadly oriented' (OECD, 1999c, p. 9), three justifications are given for this approach. First,

> ... although specific knowledge acquisition is important in school learning, the application of that knowledge in adult life depends crucially on the individual's acquisition of broader concepts and skills.

Second,

> ... a focus on curriculum content would, in an international setting, restrict attention to curriculum elements common to all, or most, countries. This would... result in an assessment that was too narrow to be of value for governments wishing to learn about the strengths and innovations in the education systems of other countries.

Third,

> ... there are broad, general skills that it is essential for students to develop. These include communication, adaptability, flexibility, problem-solving and the use of information technologies. These skills are developed across the curriculum and an assessment of them requires a cross-curricular focus.

In the PISA 2003 assessment framework (OECD, 2003b), it is noted that the assessment "is informed – but not constrained – by the common denominator of national curricula" (p. 9). In the OECD report on PISA 2000, it is stated that

The assessment is forward-looking, focusing on young people's ability to use their knowledge and skills to meet real-life challenges, rather than on the extent to which they have mastered a specific school curriculum. This orientation reflects a change in the goals and objectives of curricula themselves, which are increasingly concerned with what students can do with what they learn at school, not merely with whether they have learned it. (OECD, 2001b, p. 14)

This (identical) text also appears as a justification for the PISA approach in 2003 (OECD, 2004c, p. 20).

These extracts from the OECD reports demonstrate a fundamental shift in the purpose of the assessment. Countries are not compared on performance based on what students are supposed to have learned in school, but rather, the success of education systems is defined on the basis of the broad, real-life literacy knowledge and skills needed for personal and economic success in the future. Concerns regarding the 'fairness' of the test in terms of opportunity to learn have been replaced with a value judgement as to what knowledge and skills are relevant and desirable in current and future knowledge societies.

1.6.1.3. Differences Between PISA's Approach and Approaches in Previous Surveys

The issue of the validity of cross-country comparisons with respect to test content is not a new one, but PISA's approach puts a new spin on it and, arguably, has re-ignited the debate on the issue, which began at the time of the first international survey of educational achievement.

All international studies of mathematics and science prior to PISA have taken an approach to assessing student achievement which takes account of curricular content. Most of these studies have included measures of curricular coverage, traditionally termed measures of 'opportunity to learn' (OTL). For the purposes of international comparisons, OTL measures are significant and useful in two ways – both as a variable with which to examine and possibly explain differences in achievement, and also as a variable of interest in its own right (Floden, 2002).

As noted already, the notion of OTL has existed since the first international study (FIMS), carried out by the International Association for the Evaluation of Educational Achievement (IEA), which assessed mathematics achievement in 12 countries in 1964

(Husén, 1967a, b). According to Floden (2002) the concept has its basis in Carroll's (1963) model of school learning (which describes OTL as a continuum rather than a dichotomy, based on time learning a skill or concept). Floden also suggests the most quoted definition of OTL originates from FIMS, i.e., "whether or not... students have had the opportunity to study a particular topic or learn how to solve a particular type of problem presented by the test" (Husén, 1967a, pp. 162-163).

Analyses of curricular content have been confined to mathematics and science (e.g., Beaton et al., 1996a, b; Lapointe, Mead, & Askew, 1992; Lapointe, Mead, & Phillips, 1989) and have not been carried out in the area of reading/reading literacy. In a report on the 1991 IEA reading literacy study (the most recent international survey of reading of the school-going population in which Ireland participated prior to PISA), Elley (1992) draws our attention to constraints which must be observed when making international comparisons, namely, those regarding the student populations, the content of the tests, and the reading process. Regarding test content, he comments: "Comparisons would clearly be unfair if the measuring instrument represented the curricular emphasis of one or a few different countries" (p. 8). However, the IEA reading literacy study, rather than explicitly assessing curricular coverage in the participating countries, relied on the national submission of test materials to the international item pool to form a representative picture of the curricular aims and priorities of the countries.

PISA's orientation contrasts strongly with that of previous surveys such as TIMSS (the most recent international survey prior to PISA in which Ireland participated). For example, in the overview of the TIMSS 1999 technical report, it is noted that

> IEA studies have the central aim of measuring student achievement in school subjects, with a view to learning more about the nature and extent and the context in which it occurs. The goal is to isolate factors directly relating to student learning that can that can be manipulated through policy changes in, for example, curricular emphasis, allocation of resources, or instructional practices. (Martin & Mullis, 2000, p. 6)

Section 1.7 considers some of the problems arising from the PISA approach to assessment and reviews the efforts of countries to date to relate the PISA tests to national/regional curricula. Chapter 2 considers how the PISA tests of reading and

mathematics relate to the content and assessment of the Junior Certificate English and mathematics syllabuses, respectively.

The differences in the aims of PISA and TIMSS may be traced to the *raison d'être* of the surveys and differences in the organisations responsible for the implementation of the surveys. It was noted in Section 1.2 that the IEA was founded by a small number of individuals, the majority of whom were educational researchers and psychometricians; members of the IEA are research institutions rather than governments (Husén & Postlethwaite, 1996), hence IEA studies tend to have more of a theoretical research focus. PISA, in contrast, grew from a need expressed by OECD governments for quality indicators of educational outputs capable of capturing human capital. The research agenda therefore emphasises the economic well-being and competitiveness of countries (see also Plomp, Howie & McGaw, 2003).

1.6.2. Design

1.6.2.1. Main Characteristics

PISA surveys are conducted in schools every three years. Representative samples of schools and students are drawn to participate. In the year preceding the survey, a pilot survey is carried out in each participating country, using convenience sampling. The results of the pilot survey are used to refine and select test and questionnaire items, refine test item marking guides, and make improvements to operational procedures for the main survey. To date, there have been two survey cycles, PISA 2000 and PISA 2003; the third cycle of PISA will take place in 2006.

Key decisions regarding the survey design and how results are reported are made by the PGB, which reports to the OECD Secretariat. Each OECD country has a representative on this Board. Participating countries which are not members of the OECD can participate at meetings of the PGB as observers. The project is implemented through an international consortium of institutions: the Australian Council for Educational Research (ACER), the Netherlands National Institute of Educational Measurement (Citogroup), Westat Inc., the Educational Testing Service (ETS), and the Japanese National Institute for Educational Policy Research (NIER). The head of the consortium is the Australian Council for Educational Research (ACER), which has main responsibility for most aspects of the survey, ranging from

overseeing data operations and data entry, to data cleaning and the scaling of the test and questionnaire data. ACER has shared responsibility with the other consortium members in instrument development, sampling, field procedures and quality monitoring. International experts advise on aspects of PISA such as assessment frameworks, questionnaire and test design through 'expert groups'. There is an expert group for each subject matter and for the questionnaires expert group. There is also a technical advisory group.

The cyclical design of PISA permits the monitoring of changes in achievements and other features of the education system across time, albeit within the constraints of a cross-sectional survey design. Three domains are examined in every cycle, but the domain of focus, or 'major domain', changes with each cycle. In PISA 2000, reading literacy was the major domain, in PISA 2003, mathematics was the main focus, and in PISA 2006, science will be the major domain. In addition to reading, mathematics and science, PISA 2003 included an assessment of cross-curricular problem-solving. (It is not planned to assess problem-solving in future cycles.)

Participating students complete a two-hour pencil-and-paper test and a 30-minute questionnaire in their schools over the course of one day, while principal teachers complete a school questionnaire. The frameworks guiding the content of the tests and questionnaires are discussed in more detail in the sections which follow.

1.6.2.2. Participating Countries

In PISA 2000, 32 countries (four of these non-OECD member countries) participated. In 2002, 11 additional countries (Albania, Argentina, Bulgaria, Chile, Hong Kong-China, Indonesia, Israel, Macedonia, Peru, Romania and Thailand) administered the PISA 2000 tests and questionnaires; these are known as the 'PISA Plus' countries. In PISA 2003, all 30 OECD member countries and an additional 11 OECD partner countries participated (Table 1.1). In the OECD reports on PISA, averages are usually made with respect to the OECD average. It might be noted that the OECD average is not the same in PISA 2000 and PISA 2003, since PISA 2003 includes in addition the Slovak Republic and Turkey. The meaning of country rankings also differs across the surveys due to variation in the number of countries.

Table 1.1. Countries Participating in PISA 2000 and/or 2003

OECD Countries		*Partner Countries*
Australia	Korea (Rep. of)	Albania***
Austria	Luxembourg	Argentina***
Belgium	Mexico	Bulgaria***
Canada	Netherlands	Brazil
Czech Republic	New Zealand	Chile***
Denmark	Norway	Hong Kong-China**
Finland	Poland	Indonesia**
France	Portugal	Israel***
Germany	Slovak Republic*	Latvia
Greece	Spain	Liechtenstein
Hungary	Sweden	Macao-China*
Iceland	Switzerland	Macedonia***
Ireland	Turkey*	Peru***
Italy	United Kingdom	Romania***
Japan	United States	Russian Federation
		Serbia*
		Thailand**
		Tunisia*
		Uruguay*

*New to PISA in 2003.
**Countries administering PISA 2000 assessment in 2002 and participating in PISA 2003.
***Countries administering PISA 2000 assessment and not participating in PISA 2003.

1.6.2.3. Assessment Frameworks

This section briefly describes the assessment frameworks for PISA 2000 reading and PISA 2003 mathematics only; fuller accounts of all assessment domains of PISA 2000 and PISA 2003 can be found in OECD (2000b; 2003b). [In addition, sample tasks from PISA 2000 can be found in OECD (2001b, 2002c), and for PISA 2003 in OECD (2003b, 2004c).] Chapter 2 considers the PISA reading and mathematics frameworks and tests in more depth with reference to national curricula for English and mathematics.

1.6.2.3.1. PISA 2000 reading literacy framework

PISA does not measure whether 15-year old students are 'technically' able to read. Rather, it attempts to assess the ability of students to understand and reflect on a range of texts in various contexts likely to be encountered both inside and outside school settings. Reading literacy is defined as "understanding, using and reflecting on written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society" (OECD, 2001b, p. 21). The definition draws attention not only to comprehension process but also to higher-order reading skills. Reference to participation in society emphasises the role of reading literacy in economic, political,

cultural, occupational, and social life. In operationalising this definition, three dimensions are identified: the content or structure of texts; the reading processes that need to be performed; and the context in which knowledge and skills are applied.

The reading test included two text types (structures) – continuous and non-continuous. Continuous texts consist of sentences arranged in paragraphs. Non-continuous texts are often organised in matrix format, based on combinations of lists, and include charts and timetables. Almost two-thirds of items were based on continuous texts while the remainder were based on non-continuous texts.

Three broad categories of reading processes are also identified in the framework:

- Retrieving information (locating one or more pieces of information in a text);
- Developing an interpretation (constructing meaning and drawing inferences using information from one or more parts of the text); and
- Reflecting on and evaluating the content and form of texts (relating a text to one's experience, knowledge and ideas).

Almost half of the reading items assessed students' ability to interpret information, 29.8% assessed ability to retrieve information, and 20.6% assessed ability to reflect on and evaluate the structure and content of texts.

The third dimension, context, refers to the uses and purposes for which texts were constructed. The situations in which reading takes place, defined as how the author intended the text to be used, include: private, public, work and education. These three dimensions were brought together in a series of 48 texts and 141 tasks (items). Reporting scales based on the processes and text types have been developed (OECD, 2001b, Kirsch et al., 2002); contexts were not used as a basis for reporting results.

A variety of item formats were included in the assessment. About two-fifths of items had the traditional multiple-choice format. Complex multiple-choice item formats (5%) require students to select one alternative of a series of related 'true or false' type statements; short-response items (14%) require a word or short phrase as a response, where there may be a range of correct answers; closed-constructed response items (11%) are similar to short response items except that there is a limited range of possible correct responses, and open constructed-response items (31%) require one or more full

sentences, where there is a range of correct responses. All closed- and open constructed-response items required manual marking by trained marker using marking guides developed by the consortium. Table 1.2 shows the distribution of PISA 2000 reading items by text structure, process and item format.

Table 1.2. Distribution of Reading Literacy Items by Dimensions of the Reading Literacy Framework: PISA 2000

Dimension	*Number of Items*	*Percent of Items*
Reading Process		
Retrieving information	42	29.8
Interpreting	70	49.6
Reflecting/Evaluating	29	20.6
Total	141	100
Text Structure		
Continuous	89	63.1
Non-continuous	52	36.9
Total	141	100
Item Format		
Multiple-choice	56	39.7
Complex multiple-choice	7	5.0
Short response	20	14.2
Closed constructed response	15	10.6
Open constructed response	43	30.5
Total	141	100

Source: Shiel et al., 2001, Table 1.1; Wu, 2002, p. 27.
Note. Nine of the 141 items were dropped from the main study item pool.

1.6.2.3.2. PISA 2003 mathematical literacy framework

The PISA definition of mathematical literacy and the accompanying framework are heavily influenced by the Realistic Mathematics Education movement, which stresses the importance of solving mathematical problems in real-world settings (e.g., Freudenthal, 1973, 1981). Central to this approach is the process of mathematising, i.e., starting with a problem situated in a real-world context, organising the problem according to mathematical concepts, trimming away the reality through such processes as generalising and formalising, solving the problem, and finally making sense of the mathematical solution in terms of the original situation. The framework distinguishes between mathematical content and competencies.

Mathematics in PISA is concerned with "the capacities of students to analyse, reason, and communicate ideas effectively as they pose, formulate, solve and interpret mathematical problems in a variety of situations" (OECD, 2003b, p. 24). It is defined as:

... an individual's capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgements and to engage with mathematics in ways that meet the needs of that individual's life as a constructive, concerned and reflective citizen. (OECD, 2003b, p. 24)

Similar to the reading framework, the mathematics framework comprises three dimensions: mathematical topic areas or themes (overarching ideas), mathematical competencies, and situations. The four overarching ideas are Space & Shape, Change & Relationships, Quantity, and Uncertainty. In PISA 2003, 85 items or tasks, based on 54 units or mathematics problem contexts, were presented to students. Of these, 27.1% are categorised as Quantity, 23.5% as Space & Shape, 25.9% as Change & Relationships, and 23.5% as Uncertainty. In PISA 2000, just two of the four overarching ideas (Space & Shape, and Change & Relationships) were assessed (Table 1.3).

Table 1.3. Distribution of PISA 2003 Mathematics Items by Dimensions of the Mathematics Framework

Dimension	*Number of Items*	*Percent of Items*
Overarching Idea		
Space & Shape	20	23.5
Change & Relationships	22	25.9
Quantity	23	27.1
Uncertainty	20	23.5
Total	85	100
Competency Cluster		
Reproduction	26	30.6
Connections	40	47.1
Reflection	19	22.4
Total	85	100
Item Type		
Simple Multiple-choice	17	20.0
Complex Multiple-choice	11	12.9
Short Response Items	23	27.1
Closed Constructed Response	13	15.3
Open Constructed Response	21	24.7
Total	85	100

Source: Cosgrove et al., 2005, Table 1.2.
Note. Unlike PISA 2000 reading, none of the PISA 2003 mathematics items was dropped from the main study item pool.

The mathematics framework also describes three competency clusters: the Reproduction cluster, the Connections cluster, and the Reflection cluster. These are assumed to form a hierarchy. Reproduction items entail the use of routine, practiced skills (such as finding the average of a set of numbers); Connections problems are usually in routine contexts but require more active problem-solving, while problems in the Reflection cluster often require significant mathematisation, modelling and argumentation, and are usually in novel contexts. About 31% of items are classified as

belonging to the Reproduction cluster, 47.1% to the Connections cluster, and 22.4% to the Reflection cluster.

The framework identifies four situations in which it is believed that students encounter mathematics in their everyday lives: personal, educational/occupational, public, and scientific. Item formats are distinguished in the same manner as PISA 2000 – with about 33% of multiple-choice or complex multiple-choice items; the remainder requiring a written response (and 25% requiring an extended response). To date, achievement results have been reported on a combined scale and on four subscales corresponding to the overarching ideas.

1.6.2.4. Questionnaire Framework

The PISA questionnaire framework provides a conceptual overview of variables associated with achievement (Table 1.4). It is based on the education indicators framework of INES (OECD, 2005b). The framework organises variables along two dimensions – the level of the system (individuals, instructional settings, education providers, and the education system) and the manner in which the variables operate at each of these levels (outputs and outcomes, policy levers and contexts, and antecedents and constraints). The framework has its basis in the Second International Mathematics Survey (Travers & Westbury, 1989).

While each cell in the framework has a conceptual basis, it is unclear how the cells relate to one another. This is partly a result of the complexity of variables and relationships that potentially influence student learning outcomes. It is also partly due to PISA's cross-sectional design, which does not permit causal inferences. The sample design, which does not entail intact class sampling, does not permit the direct measurement of teacher- or class-level variables. Variables which relate to classrooms (e.g., disciplinary climate in mathematics class) and collected at the student level are aggregated to the level of the school.

Table 1.4. PISA Questionnaire Framework

	Outputs and outcomes	*Policy levers and contexts*	*Antecedents and constraints*
Individuals	Individual outcomes (e.g., achievement in reading and mathematics)	Levers and contexts relating to individuals (e.g., learning strategies and preferences; sense of belonging at school)	Antecedents and constraints relating to individuals (e.g., parental occupation, family structure, gender)
Instructional settings	Outputs and outcomes at the classroom level. Not measured in PISA.	Levers and contexts relating to the classroom (e.g., disciplinary climate in class, teacher support in class)	Antecedents and constraints relating to the classroom. Not measured in PISA.
Outputs and outcomes	Outputs and outcomes at the school level (e.g., aggregates of individual outcomes)	Levers and contexts relating to the school (e.g., school resources, admittance and grouping policies)	Antecedents and constraints relating to the school (e.g., school type, location, funding, social composition)
Education system	Outputs and outcomes at the system level (e.g., aggregates of individual outcomes, equity-related outcomes)	Levers and contexts relating to national level (e.g., system-level aggregates; other OECD data sources)	Macro-economic and demographic context (e.g., system-level aggregates; other OECD data sources)

Note. Unshaded cells indicate aspects of education systems that are addressed in the PISA School and Student Questionnaires. Shaded cells are not examined directly by PISA.
Source: OECD, 2005b, Figure 3.1.

1.6.2.5. Target Population and Sample Design

The target population for PISA is 15-year-olds enrolled in educational institutions.[6] Testing was to take place in a six-week period between March 1 and August 31 and eligible students had to be aged between 15 years three months and 16 years 2 months at the time of testing (with a one-month variation permitted).

Schools were sampled using probability proportional to size sampling, with school size based on an estimate of the number of 15-year-olds enrolled in the school. Standards relating to population coverage, sampling precision, school- and student-level exclusions, and response rates have been established. These are discussed in more detail in Section 1.7.

All national centres were required to provide a sampling frame to the PISA consortium which was to correspond to their national defined target population (i.e., the international target population minus *á priori* school-level exclusions). Centres were also asked to identify stratifying variables in order to improve sample efficiency, and to ensure complete population coverage in the sample of schools The sampling frame was

[6] In PISA 2003, this definition was restricted to grade 7 (first year) and higher (OECD, 2005b).

sent to the PISA consortium accompanied by supporting documentation and the consortium drew the school sample (Krawchuk & Rust, 2002, pp. 39-53; OECD, 2005b, pp. 46-60).

Students were sampled using KeyQuest (Volodin et al., 2003), software developed specifically for student sampling and data entry. Once participating countries received the school sample, they used KeyQuest to draw a random sample of age-eligible students from the selected schools. The number sampled per school was 35, or, in schools with 35 or fewer students, they were sampled with 100% probability (Krawchuk & Rust, 2002, pp. 53-56; OECD, 2005b, pp. 64-66).

1.6.2.6. Test Design and Scaling

In PISA 2000, reading took up 270 minutes of testing time, with 60 minutes for each of mathematics and science. Items were assigned to 11 half-hour clusters (seven of these reading, and two of each of mathematics and science). These were assembled into nine two-hour test booklets. The design was not balanced (not every cluster appeared in each of four possible booklet positions); hence estimates of item difficulty are not independent of the position in the test booklet.

In PISA 2003, mathematics took up 210 minutes of testing time, with 60 minutes for each of the three minor domain. Items were located in 13 half-hour clusters (seven mathematics and two clusters for each of the minor domains). These were rotated across 13 two-hour test booklets such that each appeared once in the four positions, giving a balanced booklet design and item difficulty estimates which are not confounded with position in the test booklet. Students were randomly assigned to booklets in both years.

In both years, a one-hour booklet was developed for use in schools for students with special educational needs in countries where more than 5% of 15-year-olds were enrolled in such schools. This was developed to reduce the level of school- and student-level exclusions (OECD, 2005b, p. 17).

In both PISA 2000 and PISA 2003, the mixed coefficiencts multinomial logit model was used to scale the achievement data (Adams, Wilson & Wang, 1997). It is a generalised form of the Rasch model. Items are described according to a fixed set of

unknown parameters while the student outcome (the latent variable) is a random effect. The model is conditional in the sense that responses are conditional upon the latent variable. Normally, this type of model requires one to assume that students have been sampled from a normal population. However, the population model can be replaced with a regression model whereby the latent variable is estimated using known values of the students (see, e.g., OECD, 2005b, pp. 119-135).

The scaling of the cognitive data in both surveys was carried out using ConQuest (Wu, Adams, & Wilson, 1997). First, items were calibrated for each country's dataset using unweighted data. This allows for the identification of items with suspicious psychometric properties that may need to be dropped from the national item pool (or if poor in many countries, from the international pool). Second, once decisions had been made regarding the treatment of items, international item parameters were set by applying the model to a pooled sample of 500 students from each OECD country. Third, student scores were generated on the basis of the international item parameters. As with all models based in item response theory, the student proficiencies are not observed – they must be inferred from the observed responses to items. Hence in PISA 2000 and 2003, five plausible values (imputed scores) were generated for each student for each achievement scale. The use of plausible values allows for the uncertainty arising from the fact that student ability is not directly observed, and that the model is probabilistic (OECD, 2005b).

The justification for the generation of the particular achievement subscales is given as follows: "Wherever multiple scales were under consideration, they arose clearly from the framework for the domain, they were seen to be meaningful and potentially useful for feedback and reporting purposes, and they needed to be defensible with respect to their measurement properties" (OECD, 2005b, p. 252; see also Turner, 2002, p. 196).

Proficiency level construction occurred in a number of stages, using an iterative process whereby steps were revisited and progressively refined. The first phase involved generating a description of proficiencies at various points on the achievement scales. This began with the identification of possible subscales. Then, a skills audit of items was carried out by members of the relevant subject expert group. In the case of partial credit items, each score level was evaluated separately. This process resulted in a

description of skills associated with different points on the achievement scales. In the second phase, the field trial data were used to derive difficulty estimates for each item. These were plotted against the student ability estimates, giving an indication of the utility of each scale from a measurement perspective. That is, the closer the match between the distribution of student ability and item difficulty, the better the utility of the scale (see, for examples, Turner, 2002, Figures 22, 23, 24, pp. 151-152). In the third phase the two steps were combined, whereby the described skills became associated with difficulty levels. This allowed the identification of clusters of skills and the possibility of describing proficiency at different regions of the scale. The descriptions were re-evaluated when the main survey item data became available. The scale descriptions were then validated, though this process is described in a vague manner, i.e., through a review by subject experts of the proficiency descriptions against material "that enabled them to judge PISA items against the described levels" (OECD, 2005b, p. 253) and through a review of participating countries of the descriptions.

The second major phase involved assigning cutpoints to the scales. As the OECD (2005b, p. 254) notes: "This is both a technical and practical matter of interpretation what it means to be at a level, and has significant consequences for reporting national and international results." Two principles were established in developing useful interpretations of what 'being at a level' means. First, skills should be considered as continua and cutpoints are essentially arbitrary. Therefore, it is only useful to regard students as having attained a particular level if this would allow for certain expectations about what students at a level are capable of in general. This was operationalised as follows: as a minimum, students needed to be more likely to get tasks at a level correct than incorrect. Specifically, the PISA proficiency scales for both PISA 2000 and PISA 2003 have the property that students with a score at the bottom of a level are expected to respond correctly to about 50% of the tasks at that level; those at the top have about an 80% chance. The second principle was that the meaning of being at a level should be consistent, regardless of the level (other than the highest and lowest, which are unbounded).

In both PISA 2000 and PISA 2003, detailed descriptions of the skills associated with each proficiency level for the combined scales and each subscale are provided, both in the OECD reports of the surveys, and in the technical reports (Kirsch et al., 2002, pp.

39-41; OECD, 2001b, p. 36; OECD, 2005b, 261-268; OECD, 2004d, pp. 47, 55, 68-69, 78-79, 85-86; Turner, 2002, pp. 204-207). In PISA 2000, there are five proficiency levels associated with the reading scale; in PISA 2003, the mathematics scale has six levels of proficiency. In both cases, a cutpoint below which PISA does not reliably assess student skills ('below Level 1') was identified.

1.7. Issues in the Interpretation of Outcomes of International Assessments

"It is axiomatic that there is no point in conducting a study just for the sake of doing it or for the sake of keeping an organisation going" (Husén & Postlethwaite, 1996, p. 139). Vitally, a match must be made between a survey design and that which education ministries in the participating countries perceive to be important. The quality of the study itself is paramount. Kellaghan (1996), Beaton et al. (1999) and Postlethwaite (1999) have identified a number of conditions and requirements of international assessments of educational achievements. I outline these here and evaluate PISA in accordance with them. (In Chapter 2, I consider PISA further in the Irish context, and explain why the particular themes have been chosen for consideration in this thesis.) For reasons of brevity, I do not evaluate PISA on every possible condition and requirement, such as the methods used to assure the quality of the procedures and data, and translation procedures. The general consensus with respect to these two aspects of PISA is that quality assurance procedures were highly satisfactory, and that the methods used to translate the materials represent an improvement from the methods used in previous surveys (e.g., Goldstein, 2004; Smithers, 2004). A technical evaluation of PISA (Hambleton et al., 2005) which considers whether the test design is optimal, whether the procedure used to link achievement across cycles, and whether there might be alternative ways to scale the data suggests that PISA may be improved in some of these respects. However, the content of that highly technical review, which focuses on IRT scaling and test design, is not the focus of the present section. [In any case, other authors, such as Blum, Goldstein, & Guérin-pace (2001), Goldstein (1995) Wu and Adams (2002), and Zabulionis (2001) have reviewed aspects of these surveys which relate to the scaling of the test data.] This evaluation of PISA focuses more on the utility and interpretability of results. Before evaluating PISA along these lines, I consider some general factors which can give rise to the misinterpretation of the results of international assessments with the aim of demonstrating that, even if a survey meets the conditions and requirements under consideration, its results are still prone to misinterpretation.

1.7.1. Factors Giving Rise to Misinterpretation of the Results of International Assessments

Several factors giving rise to the misinterpretation of international assessments have been identified. First, comparisons of the findings of different studies have been made without taking differences in the survey design, the participating countries, etc. into account (cf. O'Leary et al., 2001). Second, league tables (country rankings) have been misinterpreted; e.g., taken to be absolute measures of standards, without taking measurement error and other considerations into account (Goldstein & Spiegelhalter, 1996). Third, explanatory analyses have been misinterpreted also. Some common errors include ecological fallacy (assuming that the strength of association at one level is the same at another level of the education system), inferring causation from a cross-sectional design, failing to take inter-relatedness of background measures into account, and misinterpretation of variance components (Beaton et al., 1999; Goldstein, 1995; Kellaghan, 1996; National Academy of Education, 1993; Postlethwaite, 1995, 1999; Westbury, 1989). The media have been identified by a number of authors as a source of misinterpretation and distortion of survey results (e.g., Kellaghan, 1996, 2001; Rothman, 2002).

Perhaps at the heart of the problem lies the belief that an empirical quantitative approach somehow removes all bias in the results to the extent that cultural, economic, linguistic, education system etc. differences can be ignored. Haertel (1997) comments with respect to TIMSS:

> The rhetoric of "natural experiments" is seductive. It conjures up an image of the world as a great laboratory, with different countries trying alternative educational approaches and TIMSS as the common examination to see which approach worked best. But TIMSS is a comparative observational study, not an experiment.... the students of different nations, are not interchangeable.

A second source of the problem is the natural tendency for rankings to be interpreted competitively. For example, with respect to FIMS:

> The tests were not devised primarily in order to make total score comparisons between countries possible and certainly not as yard stick for an "international contest".... The tests are to be used primarily for comparisons between school systems both within and between countries.... Though they are considered to be of minor significance for the

project, one can hardly avoid being interested in national differences in average score... (Husén, 1967b, p. 26.)

A further source of misinterpretation is that some of the concepts associated with survey design and statistical analysis are complex to non-practitioners (e.g. policymakers and the media). Postlethwaite (1999) comments, using the example of ecological fallacy:

... it is not easy to explain the ecological fallacy to politicians, journalists, and members of the general public who tend to look towards simple uni-dimensional solutions for solving multi-factorial and multi-dimensional policy problems. More people are needed as 'information brokers' who digest many of the complex cross-national reports and then show how these can be used to make *informed decisions.* (p. 57, italics in original)

The International Reading Association's Task Force on PISA sums these difficulties up:

The challenges and pitfalls of comparing individuals, groups and nations are legion. Yet, compulsively, social scientists and psychometricians measure, gauge and scale the abilities, talents, and performances of peoples from diverse walks of life and disparate regions of the globe, all in an effort to *compare.* What is learned from these comparisons depends in no small way on how thoroughly those taking the measurements understand what makes each individual or group being measured unique, and what makes each cultural context different from others. Without these understandings, data are easily misinterpreted, and generalizations too easily oversimplified. (International Reading Association, 2003, p. 3, italics in original)

1.7.2. Survey Content

There should be evidence that the achievement measure and other data collected address the aims and purposes of the study. This section considers the validity of PISA's claims about what it measures and the interpretability and utility of the results within participating countries. It then shows how PISA's approach constrains interpretability and highlights inconsistencies in the OECD Secretariat's and PISA Governing Board's defence of PISA's approach. Then, curricular analyses undertaken by countries participating in PISA are described and it is shown that the limited nature of this research and lack of comparative framework for analysing curriculum act as barriers to some aspects of the interpretability and utility of the PISA results.

1.7.2.1. Validity of the OECD's Claims of What PISA Measures

PISA data are intended to provide an indication of 'preparedness for life' of 15-year-olds which are in turn meant to provide an indicator of human capital. Bonnet (2002),

however, has questioned the predictive validity of the PISA assessment, citing statements by the OECD such as "students who demonstrate high achievement levels are more likely to be productive workers and members of society when they leave the education system" (OECD, 1996, p. 193; cited in Bonnet, 2002, p. 388) and "In a world increasingly dominated by technology, knowledge of and skills in mathematics are central to the ability to compete in the global market place" (OECD, 1998, p. 331; cited in Bonnet, 2002, p. 388).

Bonnet argues that the relationship between a country's educational performance and its economic performance is weak (see also Coulombe, Tremblay, & Marchand, 2004, p. 9). In the absence of longitudinal measures, which could follow the outcomes of 15-year-olds over time, there is no definitive means of supporting or refuting the OECD's arguments linking the education system and economic performance, however. Goldstein (1995; 2004) also makes this point and stresses the need for longitudinal surveys. Statistics Canada is conducting a longitudinal survey, taking PISA 2000 as its starting point with follow-ups every two years (Applied Research Branch Strategic Policy, Human Resources Development Statistics Canada, 2000), but to date, the only available data relate to upper post-primary schooling and dropout (Bushnik, Barr-Telford, & Bussière, 2004); labour market outcomes/pathways of the PISA 2000 cohort are yet to be established.

Kellaghan (1996) has also criticised the claims of the OECD in this regard, pointing out that the limited achievements of sampled students, seen as measures of 'human capital', should not be equated to the human capital of a nation, first because such sampled students are not yet directly contributing to the human capital of a nation (i.e., in the workforce) and secondly, because there are many more skills that can be reasonably assumed to contribute significantly to human capital but which are not measured in international assessments (e.g., decision-making, problem-solving, team work and so on). To put it succinctly:

> While educational level [achievements] may intuitively seem to be an important factor in determining a country's economic activity, the precise relationship between educational achievement (which can be defined in many ways) and economic productivity growth rate merits further investigation. (p. 155)

Since the PISA assessment is intended to be 'forward-looking', linked to individual and societal economic success, the absence, to date, of empirical evidence affirming the predictive validity of PISA's approach must be regarded as a significant shortcoming.

Attempts have been made, however, using the IALS data, to establish the relationship between literacy levels and economic growth of countries (Coulombe, Tremblay, & Marchand, 2004). In Coulombe et al.'s study, results by age cohort were used to estimate changes in literacy levels over time; e.g., the scores of 52- to 60-year-olds in the 1995 sample were used to estimate the scores of 17- to 25-year-olds in 1960. This approach suffers from a significant methodological limitation in that it assumes that individual literacy levels are static over long periods of time (or, that if there is variation, it is not biased in one direction or another). A second shortcoming is that the analyses do not demonstrate that literacy *causes* economic growth; merely that one is associated with the other. The analyses indicated that country-level growth in literacy skills explained 55% of economic growth rates. Results also showed that gains at the highest levels of literacy were not associated with economic growth, but that a reduction in low literacy levels was. Coulombe et al. argue that measures of human capital in terms of literacy have superior predictive validity compared with years of schooling or measures of educational attainment. The study provides tentative evidence for the validity of literacy as a measure of human capital which in turn is relevant to the economic success of countries, but additional corroborating evidence using longitudinal survey designs is needed.

While it may be, for now, impossible to refute or affirm the OECD's arguments regarding student achievement and its relationship to future economic performance, a comparison of mathematics and science as assessed in TIMSS 1999, NAEP, and PISA 2000 carried out by the Educational Testing Service (ETS) in the US provides some evidence of face validity (Nohara, 2001); i.e., that PISA does measure mathematics and science skills in a manner more relevant to real-life contexts. On making this argument, I am assuming that the item ratings in this study of task characteristics such as real-life relevance, multi-step reasoning and open-constructed response formats are indicative of real-life literacy.

Regarding mathematics, Nohara (2001) found that PISA items were more likely then TIMSS items to be rated as having real-life relevance and not presented in the language of mathematics (97% compared with 44%). Multi-step reasoning was also more prevalent in the PISA items (44% compared with 31%) and PISA items were more likely to require the interpretation of graphics (91% compared with 45%). The relative emphasis in TIMSS on number sense, properties and operations (arguably lower-level skills) was much higher than in PISA (46% of items compared with 9%). The review also found that PISA items (other than number sense, properties and operations) were more likely to require computation than TIMSS items. PISA mathematics items were also rated as being more difficult than TIMSS mathematics items, taking item format, context, multi-step reasoning and computational requirements into account in difficulty ratings.

The review also found that PISA was notably higher than TIMSS in its emphasis on data analysis, statistics and probability (31% compared with 11%), and also in geometry and spatial sense (22% compared with 11%). Emphasis on measurement is also a little higher in PISA (15% compared with 11%). The percentages of items assessing algebra and functions are the same in the two assessments (19%). TIMSS mathematics items were mostly multiple-choice in format (77% compared with 34%), while comparatively more PISA mathematics items were short response in format (50% compared with 20%). About 15% of PISA items were extended response format, compared with just 4% of TIMSS items.

In the case of science, Nohara (2001) found that 66% of PISA science items were judged to be relevant to scientific contexts outside the classroom or laboratory, compared with just 16% of TIMSS items. The cross-curricular claims of PISA are evident in that 20% of science items required mathematical reasoning, compared with just 8% of TIMSS items. Further, 77% of PISA items compared with 31% of TIMSS items were rated as requiring multi-step reasoning. In assessing the overall difficulty of the assessments, the panel examined response type, context, multi-step reasoning and mathematical skill. PISA again ranked higher than TIMSS on three of these four factors (the exception being response type).

The review also found that PISA 2000 and TIMSS 1999 were comparable in the percentage of items assessing life science (30% in TIMSS and 34% in PISA), PISA placed more emphasis on Earth science (43% compared with 22%) and TIMSS placed more emphasis on the physical sciences (50% compared with 37%). Proportionately more TIMSS items were multiple-choice (73% compared with 60%), and more PISA science items required a short response (17% compared with 6%).

A comparison of PISA 2003 and TIMSS 2003 by the ETS is expected to be published in the autumn of 2005.[7]

1.7.2.2. Interpretability and Utility of PISA in Individual Countries

Kellaghan (1996) argues that a single assessment may not adequately capture the outputs of a wide variety of curricula:

> ... it is extremely unlikely that the ensuing measures accurately reflect the curricula of all countries, and indeed some aspects of curricula might not even have found a place in the measures. ... if different curricula are associated with different patterns of achievement, the interpretation of achievement differences between countries will be problematic. (p. 153)

Postlethwaite (1999) distinguishes between different approaches to the construction of the tests. Some are based on the combined intended curricula. Others are based on the views of subject matter experts irrespective of the content of curricula. Others still are based on skills/knowledge deemed necessary for coping in society. Yet others are based on employers' wish lists. The approach underlying PISA is a combination of the second and third approaches. Postlethwaite would appear to be in agreement with Kellaghan's (1996) position, since he laments that these different aspects or views of what constitutes the test domain are not included as subtests (citing time and money as the limiting factors).

Beaton et al. (1999) discuss the measurement instruments used in international assessments. Their position is that curriculum *must* be taken into account in the construction of the tests. They argue:

[7] Lemke, personal communication, July 26, 2005.

> The *test* instruments must cover the intended curriculum of the participating countries. This normally involves a two-stage process: first a content analysis of the curricula…; second, an arrival – on the basis of the first step – at the international blueprint of the tests. (p. 45, italics in original)

PISA, as we have seen, makes no attempt to assess the intended curricula of participating countries. Here, I review research in countries other than Ireland which have attempted to address the extent to which PISA matches curricular aims and assessments. I then explain why PISA's approach to assessing achievement is inadequate in the absence of supporting information about countries' curricula.

1.7.2.2.1. Why PISA's approach constrains the interpretability/utility of results

PISA's approach does not necessarily represent a solution to the question of whether the content of the assessment is equally relevant to the countries participating in the assessment:

> There is a curious contradiction in the design of PISA. It is intended to be a knowledge base for policy analysis. Yet, it explicitly rejects attempting to assess what pupils have learned in relation to the school curriculum. This puts the onus on PISA to demonstrate that non-curriculum based tests can be used to derive policy conclusions for educational systems. (Smithers, 2004, p. 38)

Smithers (2004) argues that ignoring curriculum does not eliminate it as a factor to be considered, that cross-country differences in the degree of curriculum match is a source of bias in PISA, and the lack of a curriculum match analysis severely limits PISA's explanatory power. Similarly, Prais (2003) argues that the results are "unlikely to be of specific direct help to schools, or to educational policy-makers" (p. 152). Kaiser et al. (2002) argue that curricular issues must be considered in interpreting results of international comparative studies. Goldstein (2004) has also commented on the tension between the PISA approach to assessment and the OECD's claims that it can be used to draw inferences about the performance of education systems. This 'curious contradiction' points to the need for curricular analyses to be carried out as part of the analysis and reporting of the PISA results at national level, and, ideally, at international level. The OECD has been accused of 'sidestepping' the issue of curriculum with respect to PISA (Prais, 2003). The material reviewed in the following section does indeed suggest that there has been some avoidance by the OECD Secretariat and PISA Governing Board on the possible resolution of this issue.

1.7.2.2.2. Evidence of inadequate responses to requests for enhanced interpretability and utility

In theory, the PISA approach to assessment and cross-national comparisons of curricular content, etc. are not incompatible. While comparisons of curriculum are complicated due to the existence in many countries of multiple curricula (or no curricula), or curricula which differ depending on whether students are in upper or lower post-primary education, they are not impossible, as previous efforts have shown. In particular, the approaches taken in TIMSS 2003 represent an advance from previous analyses of curriculum in TIMSS 1995 (cf. Beaton et al., 1996a; Mullis et al., 2004) whereby, for example, possible differences in academic tracks were taken into account in 2003. However, there does appear to be resistance to conducting an international comparison of curriculum with respect to PISA, even if participation in the exercise were to be on a voluntary basis. This issue did not really arise in PISA 2000, where reading literacy was the main domain, since, as noted, previous surveys of reading have tended not to examine the relationship between the international test and national curricula. With PISA 2003, however, when mathematics became the main focus, the disparity between PISA and previous surveys was much more in evidence. Hence, it was not unexpected that in May 2004, the UK Department for Education and Skills (DfES, 2004) asked the OECD Secretariat to circulate a proposal to members of the PISA Governing Board for a multilateral study (participation on a voluntary basis) which was to examine

1. classroom practices;
2. impact of 'key drivers' for system-wide educational reform; and
3. match between national curricula and assessments and the PISA instruments.

The position of the DfES on studies such as PISA is that by themselves, the studies are "…not sufficient to allow us to draw reliable conclusions about the impact of policies in different countries in recent years. We need a deeper understanding of why the distribution of outcomes which PISA revealed arose" (DfES, 2004, p. 4). The DfES proposed a 'PISA follow-up study' to address these gaps in knowledge and understanding. The first aspect of this proposal is a curriculum and assessment match, since "before we can begin to explain how such policies have influenced the relative performance of OECD countries, we must understand as fully as we can the extent to

which the PISA results relate to national curricula and the way those curricula are assessed" (p. 5). The study proposed by the DfES would not only help explain how countries differ but also yield information on the extent to which countries teach PISA skills and knowledge (as opposed to these being incidental by-products of schooling, and/or products of out-of-school learning). The proposed methods (not described in any detail) were a combination of desk research (item analysis) and interviews with nationsl curriculum experts.

The Secretariat and PISA Governing Board had the following response to this proposal:

> The Secretariat considers that country co-operation on the first two of these issues could significantly advance our understanding of the factors shaping the quality of educational outcomes but sees considerable methodological difficulties and workload difficulties with the third.... The PISA Governing Board considered the issue of curriculum match for the PISA assessments but noted the difficulties which the Third International Mathematics and Science Study (TIMSS) encountered with such work. (DfES, 2004, pp. 2-3)

The Secretariat goes on to note that the detailed curriculum analyses produced by Schmidt and his colleagues as part of TIMSS 1995 (Schmidt et al., 1997) failed to produce a conclusive description of national curricula. It further notes that curricular analyses may provide additional information about differences in content but leave questions about the (more important issue of) quality of instruction unanswered.[8]

The Secretariat concluded:

> In recognition of these difficulties, the PISA Governing Board chose not to use the common denominator of national curricula as the starting point for the PISA assessment but instead to use the knowledge and skills that countries agreed were important outcomes in each assessment domain and then to operationalise them in the PISA assessment frameworks... . (DfES, 2004, p. 3)

However, the Secretariat has acknowledged the relevance of task composition (DfES, 2004). It points out that, during the item review phase, countries were asked to rate the curricular relevance of each item, as well as the overall interest level to and relevance for 15-year-olds, relevance to PISA's assessment frameworks, and finally, to provide an

[8] However, the OECD concedes that several important aspects of the quality of schooling are not assessed by PISA in any case (see OECD, 2005b, p. 17; also Smithers, 2004, p. ii)!

overall priority for inclusion rating. The overall priority ratings for the PISA 2000 reading items were used to compare each country's performance on *all* items with only those items which were rated as having a high overall priority for inclusion. Only two countries' rankings changed significantly – Korea (ranked lower) and Norway (higher) (Adams & Wu, 2003). The Secretariat cites this research as evidence that the PISA approach to assessing achievement allows for valid comparisons of countries.

There are some conceptual and methodological problems with the analyses and how they have been cited by the Secretariat. There are also some difficulties with the position taken by the OECD Secretariat and PGB. First, the re-ranking of countries by Adams & Wu (2003) is not based on curricular relevance, but on an *overall* priority rating which takes interest, relevance, and fit to the PISA framework, as well as curricular relevance, into account. Second, although item review guidelines for this rating exercise were provided to countries, it is not known who carried out the review in each country. In some countries, the PISA National Project Manager would have completed the review; in others, a panel of curriculum experts. In some countries, one person only may have done the review; in others, the views of a number of individuals may have been combined. So in this sense, one is not comparing like with like. Third, depending on the political, educational, and other agenda of countries completing the review, the reasons for assigning ratings of curricular relevance and overall priority are also likely to differ. Fourth, analyses were carried out with respect to reading items only. The extent to which curricular content will affect performance on a test may vary according to the subject domain assessed and in fact is likely to be more critical in the case of mathematics and science than in the case of reading. Fifth, adjustments were not made for the relative difficulty of the subset of items which were given high priority ratings by each country. For example, if one country were to give a particularly difficult subset of items high overall priority for inclusion ratings, while another were to give a particular easy subset of items, then the re-ranking of these two countries is confounded with the relative difficulty of the subset of items chosen. Moreover, one might view the position of the OECD Secretariat as overly defensive. It assumes that the analyses proposed by the DfES are necessarily costly and as complex as those associated with TIMSS 1995.

It appears, therefore, that despite attempts by the PISA Governing Board and OECD Secretariat to defend and justify PISA's approach to measuring skills, anomalies still remain.

1.7.2.2.3. Research on PISA and the curriculum in countries other than Ireland

In response to an email sent by me to PISA National Project Managers on June 8, 2004, informing countries about Ireland's planned curriculum analyses for the 2003 national report, and requesting information about what countries were doing, if anything, to compare national curricula with PISA, I received affirmative replies from ten countries.

The USA PISA representative (Lemke, personal communication, June 9, 2004) indicated that while the USA does not have a national curriculum as such, comparisons between PISA 2000, TIMSS 1999, and performance on the USA's national assessment programme (NAEP), have been carried out by the National Center for Education Statistics (NCES), the outcomes of which were reviewed in Section 1.7.1 (Nohara, 2001). The Slovak Republic has analysed the PISA mathematics items with respect to their national curriculum, but results have not yet been published (Korsnakova, personal communication, June 11, 2004). The Italian PISA team indicated that researchers in Italy were contemplating a rating of the PISA 2003 test items but had concerns about the reliability of the rating methods (Siniscalso, personal communication, June 13, 2004). The PISA team in Belgium (French Community) has carried out some analyses which rated each mathematics item in terms of the competencies described in the national curriculum and identified the grade and track which corresponds to the competency. These results have also been used to prepare a document aimed at teachers which presents some PISA mathematics items and describes where the competencies lie in relation to the national curriculum (Demonty, personal communication, June 14, 2004). Switzerland has also analysed the PISA mathematics assessment with references to curricula in the 26 cantons (Zahner, personal communication, June 15, 2004). Switzerland also published a report on PISA 2000 which examined the extent to which PISA skills correspond to curricula for reading, mathematics and science (Zahner, personal communication, June 15, 2004). Poland carried out an analysis of PISA 2003 mathematics items with the aim of detecting weak and strong points of students' competencies (Federowicz, personal communication, June 15, 2004); however this seems to have been more of an analysis of student performance at the item level than a

comparison of PISA with the curriculum (Federowicz, personal communication, July 27, 2005). In Germany a group of curriculum experts have rated the extent to which items fit with curricula. The analysis is similar to the curriculum analysis in PISA 2000, where the curriculum fit of TIMSS and PISA (international and national additional tests) were compared (Ramm, personal communication, June 16, 2004). Results for PISA 2003 are expected in November 2005 (Ramm, personal communication, July 26, 2005). Portugal has rated each mathematics item from PISA 2003 on a 5-point scale, ranging from 1 (mathematics ideas involved in the item are not present on the grade 9 curriculum) to 5 (mathematics ideas involved in the item are totally covered on the grade 9 curriculum) and obtained an item average rating of 4.4 (Ramalho, personal communications, June 17, 2004; July 26, 2005). Uruguay has published a comparison of PISA 2003 mathematics with their national curriculum through a classification of elements of the PISA framework in terms of the primary and post-primary mathematics curricula (Ravela, personal communication, June 17, 2004; July 27, 2005). The Danish PISA team also expressed an interest in the analyses but have not yet carried out or published analyses of PISA and the Danish curricula (Lindenkov, personal communication, July 27, 2004).

To summarise, it would appear that only a small number of participating countries have undertaken analyses concerning national curricula and PISA. (Of course, it is possible that some countries which did not respond to the enquiry may have analysed PISA with respect to national curricula.) The results of these analyses have not been widely disseminated outside the country in question. Further, comparisons of the results are hampered by a lack of international framework for curriculum analysis. The lack of dissemination and lack of comparability of methods act as barriers in discussions about the issue of the curriculum and its relation to PISA on an international level. Therefore, the framework and results of the test-curriculum rating project in Ireland, described in Chapter 2, and any conclusions about the expected familiarity of Irish students with the PISA tests, may only be described and interpreted in national context.

1.7.3. Interpretation of Proficiency Levels

In the US, the National Academy of Education (NAE) (1993) conducted a review of the procedures for setting proficiency levels and problems with their interpretation. Many of these issues also apply to international assessments. The review in fact resulted in a

recommendation to discontinue the reporting of results in terms of achievement levels. The NAE argued that the descriptions of the levels and exemplar tasks were not sufficient to illustrate the requirements associated with each level, recommending instead the use of percentile scores. Its longer-term recommendations included the development of content and performance standards as an iterative process which take changes in national standards into account.

The methods used to develop proficiency levels in PISA reviewed in Section 1.6 would appear to accord with the iterative nature of the process recommended by the NAE. However, even though it is stated in the technical reports for PISA that the cutpoints of these levels are essentially arbitrary and that the levels are to be interpreted as a continuum of skills rather than discrete categories (OECD, 2005b; Turner, 2002), this is not stated in the main reports. [Blum, Goldstein and Guérin-Pace (2001) and Kellaghan (2001) have also discussed the arbitrary nature of cutpoints. Blum et al. have shown, for example, that a small shift in the cutpoints associated with IALS dramatically lowers the estimated percentage of French adults at proficiency Level 1.] Unless one is inclined to read both the technical documentation and qualitative descriptors associated with the scales, interpreting them is problematic. This makes the distinction between at Level 1 and below Level 1 difficult to grasp. This distinction is important, since one can say with some confidence what students at Level 1 can do, but there are no exemplar items below Level 1 to allow interpretation of the performance of these students. In reporting the results, the categories at and below Level 1 are usually combined (using phrases such as 'fail to reach Level 2') (e.g., OECD, 2001b, pp. 47-49; OECD, 2004c, pp. 51, 56, 69, 74). However, in some of the media reports of the results on PISA in Ireland discussed in Section 1.8., Level 1 and below is considered as the indicator of or benchmark for low literacy; in others, below Level 1 only is considered. Whether or not these two categories should be combined is debatable, but so far, there has been no discussion of the issue. The lack of sufficient numbers of exemplar items is likely to be a problem at both extremes of the item difficulty distribution. These problems are further compounded by oversimplification and distortion by the media (e.g., Kellaghan, 2001).

The International Reading Association (2003) has noted that public examinations in participating countries may not relate well to international benchmarks implied in

proficiency levels in studies such as PISA. Given PISA's curriculum-free approach, and the fact that students are spread across several grade levels, pinning international benchmarks to achievements on national assessments or examinations at a particular point in the system is difficult. This weakens the OECD's claims that PISA can be used to establish benchmarks for educational improvement, or understand relative strengths and weaknesses of educational systems. The lack of comparative data on how the PISA tests relate to national curricula further compounds the problem. Chapter 2 reviews the efforts to link performance on PISA with public examinations in Ireland to date, and shows how the analyses in Chapter 4 attempts to address some of the problems in interpreting PISA with respect to achievements on public examinations in Ireland.

1.7.4. Target Population and Sample Design

The design of a study should match its aims and purposes. The choice of design is, Postlethwaite (1999) holds, "an educational and political problem" (p. 19). Particular considerations are the choice of the population (its age/point in the system), and whether grades or particular age groups are to be sampled. It also entails a consideration of the policy questions and which sample design might best match them. Logistic and especially financial constraints put limits on aspects of the sample design.

Survey sample designs of international assessments since the 1980s have generally entailed two-stage stratified sampling with the probability of sampling proportional to size. If the target population is based on age, the sample is usually drawn using simple random sampling of a fixed number of age-eligible students from each school. If grade-based, the sample usually entails drawing one or more intact classes from the specified grade level(s). Both approaches, however, can give rise to problems when it comes to interpreting achievement outcomes. Results need to be interpreted with respect to what the target population is, and whether the sample adequately represents it. In the case of a sample which is based on students' age, if age-eligible students are dispersed across multiple grade levels, the sample cannot be said to be representative of a particular grade level or point in the system, unless a majority of age-eligible students are clustered within a particular grade level. In the case of a target population defined on the basis of grade level, if just one intact class per school is selected, then the sample may not be representative of the target population, particularly when students are clustered within classes on the basis of achievement. If all intact classes at a particular grade level

are selected, or a random sample across all intact classes at that grade level drawn, then that sample should be satisfactorily representative. This dissertation uses the shorthand 'grade-based sample' to refer to a sample of one or two intact classes per school (i.e., the TIMSS sample design) and 'age-based sample' to refer to a random within-school sample of age-eligible students (the PISA design).

An important issue to consider with respect to the sample design is the manner in which variance in achievement is partitioned between and within schools. Postlethwaite (1995) has noted that, in general, samples with low between-cluster variance may be regarded as more homogenous in terms of achievement, while those with high between-cluster variance are more heterogenous, and schools differ more in terms of achievement. The between-cluster variance is expressed as a proportion, a percentage, or a correlation (*rho*, the intra-class correlation). The OECD (2001b, 2004c) has taken the interpretation of the between-school variance statistic a step further with respect to PISA, taking low between-school variance as an indication of equity of a system. For example, for countries with comparatively low between-school variance and average or above-average performance on PISA 2003, the OECD (2004c, p. 163) claims that "parents in these countries [including Ireland] can be less concerned about school choice in order to enhance their child's performance, and can be confident about high and consistent performance standards across schools in the entire education system." The OECD (2001b, 2004c) does acknowledge, however, that the interpretation of between-school variance in achievement should take multiple features of education systems into account, including socioeconomic composition, sub-national differences, public/private management, parental choice, policies relating to ability tracking, the selection procedures of schools, and curriculum differentiation (OECD, 2001b, pp. 62-63; OECD, 2004c, p. 163).

At least two significant problems with the OECD's interpretation remain, however. First, although international survey experts have stated that a consideration of the interpretation of between-school variance needs to be made with reference to the sample design (e.g., Postlethwaite, 1999), this is not flagged in OECD reports on PISA, and may result in misinterpretation of the results, particularly if one is attempting to compare results on PISA with a survey which uses intact-class sampling such as TIMSS. I explore this issue further in this section through a comparison of the two types

of sample and explain the likely consequences for the interpretation of between-school variance for each.

Grade-based samples such as used in TIMSS are designed to allow an analysis of the knowledge and skills of students at a particular point in the system, and this can yield potentially useful policy information, particularly if the grade level corresponds to an end-point of a study programme (such as third year, which corresponds to the end of the Junior Cycle). However, there are problems with this type of design when drawing comparisons cross-nationally. Students enter the system at different ages (in OECD countries, this ranges from three to seven years; OECD, 1999b), retention/repetition rates differ (although quantitative data on this is lacking[9]; see also O'Leary, 2001, pp. 196-197), and progression from one level of the system to another occurs at different points depending on the country (Goldstein, 1995, 1997; OECD, 1999b). Further, as already implied above, there are problems with a grade-based sample design if class allocation is made on the basis of student ability and if only one or two intact classes per school are drawn (as in TIMSS). If this is the case, then estimates of between-'school' variance – used as an indicator of the homogeneity of education systems – may be confounded with student ability (and hence inflated) (see Postlethwaite, 1999). A corollary of this is that explanatory models of achievement (e.g. Martin et al., 2000b, pp. 71-98) confound between-class and between-school variance, particularly if only one intact classroom per school is used in the analyses.

Age-based sample designs are used when the aim is to examine how education systems have educated an age cohort, as with PISA. However, while the PISA sample design is more efficient than TIMSS in that student-level estimates are more precise, students of a similar age may be spread across a number of grades which may correspond to differing study programmes (O'Leary, 2001; Postlethwaite, 1999; Smithers, 2004). Hence, a consequence of the PISA sample design is that performance is not anchored to a particular point in the system (unless grade levels are defined on the basis of age), and this renders comparisons of performance with reference to the structure of systems and the content and delivery of national curriculum and assessment problematic. The

[9] The only international comparative data on grade repetition that I was able to find comes from the UNESCO educational statistics database and the data are available at primary level for just 11 of the 28 participating OECD countries, and not available at all for second level (http://132.204.2.104/unesco/eng/TableViewer/Wdsview/dispviewp.asp?ReportId=52).

interpretation of results with reference to the dispersion of students across multiple grades/programmes is further compounded by PISA's 'curriculum-free' approach to assessment. Prais (2003) raises this issue in his critique of PISA, and the age-based approach is defended in a rejoinder to this critique by Adams (2003), but Smithers (2004) interprets Adams' response to Prais on this issue as further evidence of the difficulties in making comparisons across countries. However, on the plus side, since the within-school sample is taken at random, estimates of between-school variance are not confounded with class allocation, and hence the interpretation of explanatory models of achievement differences between 'schools' are also more straightforward than with a grade-based sample such as TIMSS. Nonetheless, in school systems where streaming or tracking operates within schools, estimates of between-school variance from age-based samples may disguise large within-school differences in achievement. The key issue here is that conclusions about educational 'equity' can depend very much on the unit upon which the sample is based. A comparison of the variance components associated with the Irish mathematics achievement data for PISA 2000 and TIMSS 1995 in Chapter 5 further illustrates difficulties relating to the interpretation of between-school variance.

In the case of IAEP II (which had an age-based population definition and random within-school sampling, similar to PISA), O'Leary (2001) has criticised the reports for failing to highlight the issue of the distribution of students across grade levels in the interpretation of results. The same may be said of the PISA reports. He concludes:

> Above all, an effort should be made to develop procedures that allow for the outcomes of international tests to be adjusted for age, grade, and/or policies relating to repetition and social promotion. ...Given what we now know about the factors that have impinged on performance in past international studies, it seems evident that those same factors should be the focus of very close attention in PISA. (p. 198)

While some of these issues may be explored through the PISA international datasets which are publicly available, the fact that they were not highlighted in reports as issues affecting interpretation is unfortunate and leaves us in the same situation as we were with the IAEP II reports.

In discussions of between-school variance, a second aspect which the OECD does not take into account is the likelihood that the PISA measures differ across participating systems in terms of their similarity to, or difference from, the content of national curricula. There is anecdotal evidence, regarding mathematics, that countries differ with respect to the extent to which the Realistic Mathematics Education approach (e.g., de Lange, 1994, 1998) has been integrated into national curricula. As mentioned in Section 1.6, the assessment of mathematics in PISA is heavily influenced by Realistic Mathematics and processes one might associate with of horizontal mathematisation. Oldham (2001, 2002) has pointed out that at post-primary level, modern mathematics (which treats mathematics as the study of structures, rigour and logic; and vertical mathematisation, whereby high-level reasoning is applied in narrow, abstract contexts) was adopted in Ireland to a greater degree than in many other countries. She argues that the textbooks and examination papers still "reflect the focus on precise terminology and abstraction that is characteristic of the [modern mathematics] movement" (2002, p. 43), a situation which stands in strong contrast to PISA mathematics. Oldham (2001) also draws our attention to the fact that, in TIMSS 1995, teachers of mathematics in Ireland gave the lowest priority rating to understanding how mathematics is used in the real world, and comparatively high priority to the memorisation of formulae and procedures.

As noted previously, the teaching and learning of reading may be less prone to curriculum variation than mathematics or science (Beaton et al., 1999; Postethwaite, 1999). The distinction between 'school-dependent' and 'school-independent' subject areas is relevant. Kellaghan, Madaus, and Rakow (1979) have argued that subject areas which are more likely to be encountered (almost) solely in the context of school instruction, such as mathematics, may be considered school-dependent, while subject areas which may be encountered outside school instruction, and/or whose concepts and skills may be applied in a variety of subject areas (such as reading), may be considered school-independent. Thus, in a general sense, there may be more of a need to consider the contents of mathematics curricula compared with reading in interpreting survey results. The PIRLS 2001 Encyclopaedia (Mullis, Martin, Kennedy, & Flaherty, 2002) provides a description of the education systems of countries participating in PIRLS, as well as a description of teacher and teacher education, reading curriculum and instruction, literacy programmes, and assessment. The descriptions are impressionistic rather than definitive for the most part. However, many commonalities are evident

across countries, such as the integration of teaching reading into the teaching of the language of instruction; the presence of instructional goals which differentiate aural, oral, reading, and writing skills; and a distinction between various reading processes and purposes or contexts. Countries appear to vary somewhat with respect to the relative emphasis placed on literary texts; when formal reading instruction begins; the explicitness of instructional goals and targets; policies on second-language instruction; usage of textbooks; and the extent to which instruction of reading skills is integrated across the curriculum.

I further demonstrate in Chapter 2 that, in Ireland at least, the interpretation of between-school variance should be made with respect to both the subject domain assessed and whether it is intended to be curriculum-sensitive or not.

Two more issues may be raised with respect to the PISA target population which are not central to this dissertation but nonetheless worth mentioning. First, the age 15 was chosen since it is argued that this is the modal age at which compulsory schooling ends across OECD countries (OECD, 2001b). This is, arguably, however, not the most appropriate or relevant age. The compulsory school age is 16 in Canada, Denmark, Finland, France, Hungary, Iceland, New Zealand, Norway, Spain, Sweden and the United Kingdom, and 17 or 18 in Belgium, Germany, the Netherlands and the United States (OECD, 2001a, Table C1.2). In fact, the modal age is *not* 15 for OECD countries, either before *or* after a recent change in the Irish legislation relating to school-leaving[10], but rather, 16 (compare OECD, 2001a, Table C.1.2 to OECD, 2000a, Table C.1.2). Thus the PISA target population might be better described as the modal age at which students are nearing the end of compulsory schooling.

Second, countries vary with respect to the proportion of the target population that is enrolled in schools. For example, in PISA 2003, in 28 countries, the definition of 15-year-old and school-going 15-year-old are fairly synonymous, with an enrolment rate of 95% or higher; in others, such as Brazil, Indonesia, Mexico, Turkey and Uruguay, the enrolment rate of 15-year-olds is much lower (between 54% and 74%) (see Cosgrove et

[10] The minimum school leaving age in Ireland was raised from 15 to 16, or the completion of three years of post-primary education (third year/grade 9), whichever is the later, with the introduction of the Education (Welfare) Act on July 5th, 2000.

al., 2005, p. 49). There is no obvious solution to this problem other than documenting enrolment rates, which is the case with PISA. However, the data on which population estimates are taken may not be reliable in all cases. Some countries report precisely 100% enrolment (e.g., in PISA 2003, this applies to Finland, Luxembourg, the Netherlands, Tunisia and the USA; see OECD, 2005b, pp. 168-169) and are evidently using the sampling frame as a basis for population estimates.

Postlethwaite (1999) appears to be pessimistic about the possibility of attaining comparability of samples on the basis of differences discussed here. Using the TIMSS 1995 sampling information (mean age, grade tested, response and exclusion rates) as an example, he comments:

> ... it is extremely difficult, if not impossible, to arrive at exactly comparable defined target populations but the question must be whether they are reasonable comparable. (p. 27)

However what is 'reasonably comparable' is not defined by Postlethwaite. There seems no definitive solutions to problems of comparability except to accurately document between-system differences in starting ages, grade repetition, ages of transfer, grades at which national assessments or public examinations are taken, enrolment rates, and the broad content of study programmes of participating students (especially for school-dependent subject areas), and take these into account when making comparisons. The OECD has not to date provided sufficient (or sufficiently accurate) data relating to all of these aspects of participating countries' education systems to allow readers of its reports to be confident about what might be 'reasonably compared'. Documentation addressing all of these issues would be highly desirable. It would also be desirable for both the OECD and the IEA to draw more attention in their reports to the differences between the PISA sample design and that of TIMSS, and spell out some of the potential consequences regarding the partitioning of variance components.

1.7.5. Sampling Standards: Coverage, Precision and Sources of Bias

It is generally agreed that five criteria relating to sampling outcomes must be met if estimates based on a sample are to be accepted as being representative of a population

(Kish, 1965; Cochran, 1977[11]): adequate coverage of the population of interest in the sampling frame; adequate numbers of individuals sampled to provide precise population estimates; adequate statistical procedures to adjust for the clustered nature of the sample design; adequate response rates for the units or individuals sampled; and adequate statistical procedures to adjust for non-response at the cluster and individual levels. Each of these criteria is described briefly in the sections that follow, and PISA is evaluated with respect to them. It will be shown that adjustments for non-response in particular present problems to the interpretation of the results.

1.7.5.1. Sampling Standards in PISA

1.7.5.1.1. Population coverage

In PISA, schools may be excluded from the sampling frame for practical reasons (e.g., extremely small size; geographic inaccessibility) or for political reasons (e.g., language group). Students may also be excluded from the assessment (or deemed exempt from the assessment) due to special educational needs, physical disability or limited familiarity with the language of the assessment. Together, these exclusions must not exceed 5% of the target population. This is more stringent than the TIMSS 1995 criterion of 10% (Martin & Mullis, 1996) and consistent with recommendations of Beaton et al. (1999) and Postlethwaite (1999). Also, unlike TIMSS, the portion of permitted exclusions was broken down into school and student components: school-level exclusions were not to exceed 2.5% of the target population of students and student-level exclusions were not to exceed 2.5% (Krawchuk & Rust, 2002, pp. 40-41). An overall index of population coverage was computed for all countries in PISA 2000 and PISA 2003 as one indicator of the quality of the data.

1.7.5.1.2. Sampling precision

Similar to TIMSS, the PISA sampling standards required countries to sample a minimum of 150 schools and, within each school, a minimum of 35 students was to be selected, to obtain an overall desired sample size of 4500 students (assuming a within-school response rate of around 85%). If some schools on the sampling frame had less than 35 eligible students, the number of schools to be sampled had to be increased to yield the same overall sample size. The reason that this large number was required is

[11] Discussions of the issues can also be found in most technical documentation accompanying such surveys, e.g., Adams and Wu (2002); OECD (2005b).

because the clustered nature of the sample design provides less accurate achievement estimates than would a simple random sample of students. The number of schools/students ensured an achieved sample that was equivalent to at least 400 students sampled at random (the so-called 'effective sample size'), which is consistent with general sampling precision requirements (e.g., Kish, 1965). An effective sample size of 400 yields approximately the following 95% confidence limits for means, percentages and correlation coefficients, respectively: ± 0.1 standard deviations, ± 5%, ± 0.1. Further, a school sample size of 150 is recommended in order to provide estimates of school-level variables yields 95% confidence limits of around ± 16% of their standard deviations (Foy, Rust, & Schleicher, 1996). Again, this is consistent with standards discussed elsewhere (Postlethwaite, 1999).

To reduce variance in achievement estimates arising from the sample design, participating countries were asked to identify a number of so-called 'implicit stratifying' variables, or school characteristics which explain variance in achievement outcomes of schools (see Krawchuk & Rust, 2002).

1.7.5.1.3. Statistical procedures to adjust for the clustered design

As already noted, the PISA sample is based on a two-stage stratified design (i.e., schools are selected first, then students). However, the fact that students within a school are more like each other than students in different schools affects the precision of achievement estimates. If statistical analyses do not take the clustered design into account, the standard errors will be under-estimated. In PISA, the clustered nature of the design was addressed by using a specific 'bootstrapping' method, whereby the statistic of interest (e.g., mean achievement score) is repeatedly calculated for subsets of the participating schools, and the variance in these estimates incorporated into the standard error of the achievement estimate. The particular variant of the method is known as Balanced Repeated Replication (BRR), using Fay's method (Rust & Krawchuk, 2002). All analyses, whether reported in the international reports (e.g., OECD, 2001b; 2004c, d), or the national report (e.g., Shiel et al., 2001; Cosgrove et al., 2005), employed this method using specialised software called WesVar (Westat, 2000). WesVar can also incorporate the additional error introduced by the five plausible values associated with achievement estimates in the computation of mean scores.

1.7.3.1.4. Response rates

Beaton et al. (1999) have suggested a minimum response rate at the school level of 90%, and 80% at the level of the student. The PISA sampling quality standards (see Krawchuk & Rust, 2002 for PISA 2000 standards; the same apply to PISA 2003) state that a minimum of 85% of sampled schools, and 80% of sampled students, must participate. While there is a procedure for using replacements for schools that decline to participate, the initial response rate must exceed 65%. To identify replacement schools, prior to sampling, schools are sorted by explicit stratum and, within strata, by the implicit stratifying variables. Schools immediately preceding each sampled school are the first replacement schools; those immediately following each selected school are the second replacements. The first replacement school is invited to participate in the event that the original sampled school cannot participate, and the second is invited to participate if the first replacement cannot. The lower the initial response rate, the higher the final response rate with replacement schools must be. For example, at the school level, a sample with an initial participation rate of 65% is deemed representative only if the final rate equals or exceeds 95%.

1.7.3.1.5. Statistical adjustments for non-response

Non-response can occur for a number of reasons, the precise nature of which is difficult to describe and quantify. The lack of agreed-on procedures to estimate biases arising from these, and procedures which will reduce or eliminate this bias, has received increasing attention in recent years. In analyses that compared outcomes on IAEP 2 and TIMSS, O'Leary, Madaus, Kellaghan and Beaton (2000) comment that "...a particularly vexing question in international assessments (or any large-scale assessment for that matter) is the extent to which exclusions and participation rates affect overall performance." In a seminar of the UK Department for Education and Science (DfES; March 21, 2005), John Micklewright presented a paper on the topic of student and school response bias in PISA and TIMSS. He noted that the PISA response rate criteria are somewhat arbitrary and suggested that the 85% rate seemed to have been set so that, whatever the levels of non-response bias, it should not be higher than the sampling error. It was also noted that the OECD had no published criteria for the quality of bias analyses such as those included in the PISA 2000 technical report (for New Zealand, for example; Adams, Rust, & Monseur, 2002) and what was an acceptable test of bias. This section describes the statistical adjustments made to account for non-response and

explores the concept of bias and its consequences for interpreting achievement outcomes.

1.7.3.1.5.1. School-level adjustments

Adjustments for non-response in PISA 2000 and PISA 2003 were made at the school level by grouping schools into similar implicit stratum groupings (collapsing adjacent groupings where the number of schools is less than six) and applying the reciprocal of the response rate to each grouping (e.g., Rust & Krawchuk, 2002, pp. 91-93). Although this approach accords different adjustments according to school sector and gender composition of the student body in Ireland (for example), these schools also differ in other ways, some of these are known (e.g., whether the school is designated disadvantaged), and some are unknown (e.g., the morale of the teaching staff; the extent of links between the school and parents).

In research using simulated datasets, Monseur and Wu (2002) have compared the extent of bias[12] in school non-response and the efficiency of the non-response adjustment of the type used in PISA 2000, varying the proportion of between-stratum and between-school variance. They concluded that the efficiency of the non-response adjustment to control the bias introduced by non-response is proportional to the size of the between-school variance. If the between-school variance is zero, then, no matter how strong the correlation between student ability and propensity to participate, the school non-response adjustment will not introduce any bias into the estimates; i.e., it will be efficient. If, however, the between-school variance is high, the school non-response adjustment will not be efficient in controlling the bias introduced to the estimates.

1.7.3.1.5.2. Student-level adjustments

Before describing the procedures used to adjust for student non-response in PISA, it is useful to distinguish between several categories of students who do not respond in an assessment situation such as PISA. The first category includes students who are exempt from the assessment due to special educational needs. In PISA 2000 and PISA 2003, four categories of such students were identified: students with a physical disability that

[12] Here, I use the term 'bias' in the sense suggested by Monseur and Wu (2002), i.e., to indicate an over-estimation of student achievement and an under-estimation of the variation in achievement, assuming a positive relationship between student proficiency and propensity to participate.

would prevent them from participating in the assessment, students with a general learning disability, students with limited knowledge of the language of the assessment (typically, less than one year's instruction in that language), and an optional, nationally-defined exemption criterion. Table 1.5 shows detailed definitions of these exemption categories (as used in Ireland).

Table 1.5. Criteria for Exempting Students from Participation in PISA 2000 and PISA 2003 (as Adapted for Use in Ireland)

Category	*Criteria*
1	Functionally disabled students: i.e., students who are permanently physically disabled so that they cannot perform in the assessment situation. Functionally disabled students who **can** respond to the assessment should be included in the assessment.
2	Students with learning disabilities: i.e., students who are considered in the professional opinion of the school principal or other qualified staff member to be learning disabled (slow learners), or who have been identified as such following an appropriate psychological assessment. This category includes students who are emotionally or mentally unable to follow the general instructions of the assessment.
3	Students with limited proficiency in the assessment language: i.e. students who are unable to understand or speak the language of the assessment (English) and would be unable to overcome the language barrier in the assessment situation. Typically, a student who has **less than one year** of learning the language of the assessment should be excluded.
4	Students with dyslexia: i.e., students who are considered in the professional opinion of the school principal or other qualified staff member to be dyslexic or who have been identified as such following an appropriate psychological assessment. Students with mild reading difficulties **should** be included in the assessment.

A second category of non-response is students who have left the school since the list of students for that school was drawn up. Age-ineligible students, i.e., those not born in 1984 (PISA 2000) or 1987 (PISA 2003) who were included in the sampling frame in error, are similar to those students who have left the school since the list of students was drawn up, since they were included on that list in error. A third category of non-response is student or parental refusal. A fourth category of non-response is (eligible) student absence on the day of the assessment for no specified reason.

One can further categorise these groups into two broad types: students who were sampled that were either exempt or ineligible, and students that are eligible, but who did not participate in the assessment. In terms of the calculation of response rates and computation of sampling weights, the first category is not defined as non-response, but rather, as exclusion, but included in estimates of population coverage (e.g., Monseur,

Rust & Krawchuk, 2002, pp. 135-136). It is the second group, students who refused to participate, and those that were absent for no specified reason, that are regarded as 'true' non-responders for the purposes of PISA. Hence, in making adjustments for non-response, the former category is not included in the calculation (i.e., in weighting the responses of the participating students, exempt and ineligible students are not counted into the number of students that should have participated), while the latter is.

The weight adjustment used to control for non-responding students in PISA is somewhat similar to that for the school level: it assigns the average score of participating students in a given school to a student in that school who is eligible but absent on the day of the assessment. That is, the student non-response adjustment equals the ratio of the number of selected, eligible students to the number of participating students, except in very small schools with less than 15 participating students, where schools were collapsed and treated as a single unit. Schools with a student response rate of 25% to 50% were also collapsed with other schools since a school with a response rate in this interval would have a disproportionately high adjustment factor of between 2.0 and 4.0. Schools with a student response rate of less than 25% were not deemed representative and their data were removed from the database in both PISA 2000 and PISA 2003 (see Rust & Krawchuk, 2002, p. 94).

This procedure, however, is problematic since a non-participating student is regarded as equivalent to one that is present, and thus, unlike the adjustment for non-response at the school level, no account is taken of any student characteristics, such as gender, educational programme, grade level, or date of birth. This demographic information is readily available from the student tracking forms (administrative documents used to track and verify student demographics and participation status in PISA). Needless to say, other background variables, such as socioeconomic status, are not taken into account in making non-response adjustments either. It is surprising that information from the student tracking form was not incorporated into the student non-response adjustment, and also that the minimum student response rate (80%) is lower than the minimum school response rate (85%).

I take up this issue in more detail in Chapter 2, where the response rates for Ireland are evaluated with respect to biases in mean achievement and achievement variance, given

the pattern of between-school variance observed in the Irish data for PISA 2000 and 2003.

1.7.6. Issues in the Interpretation of Explanatory Models of Achievement

The use of multilevel explanatory models within the field of education beginning in the 1980s may be traced back to Coleman et al.'s (1966) work (Raudenbush & Willms, 1995; Smyth, 1999), and forms the basis of school effectiveness research, the basic premise of which is that, after taking student background factors into account, there remain differences among schools which may be ascribed to the quality of schooling (Goldstein, 1997).

Raudenbush and Willms (1995) distinguish between two conceptualisations of a school effect. The first refers to the effect of a particular school policy or practice on a student outcome. The second refers to the extent to which attending a particular school (or school with a particular characteristic or set of characteristics) modifies a student's outcome. Multilevel modelling can examine school effects in either sense.

A second major phenomenon which has been widely examined in multilevel models of achievement is the 'social context effect', whereby the social intake (average social background of the students in the school) exert an influence on achievement outcomes over and above individual student social background (e.g., Raudenbush & Willms, 1995; Willms, 2002). Research over the past three decades has shown that when children are segregated, either between schools or between classes within schools, children from advantaged backgrounds do better, and those from disadvantaged backgrounds do worse (e.g., Brookover et al., 1978; Gamoran, 1992; Henderson, Mieskowski, & Suvageau, 1978). This effect, however, has not always been detected (see Nash, 2003). That contextual effects may be stronger for low-SES students leads Willms (2002) to dub it the 'double jeopardy' hypothesis. Some of these studies suggest that the social context effect tends to be slightly larger for males than for females, and for minority status students compared with majority status students; in other words, there is a cross-level interaction between gender or other individual characteristic and the social context.

One might ask if the issues of social context effect and school effects are relevant to a consideration of PISA. The principal reason that they are is the high emphasis which they have received in international and national reports of student achievement; further, their interpretation has not been subject to a critical review in Ireland to date.

Since TIMSS 1995, the use of multilevel modelling techniques has become part and parcel of international survey reports. The multilevel models reported usually accord central importance to the variance in achievement between and within schools that is attributable to social background; they often factor out this variance in order to examine which school/class-level variables (if any) explain achievement variance over and above social background. In the PISA 2003 international report, for example, an entire chapter is devoted to the issue of student social background; a second chapter examines the extent to which, after adjusting for student social background and school social intake, school-level variables impact on student achievement (OECD, 2004c). Citing high correlations between student performance on the assessment domains, the OECD argues against the usefulness of producing explanatory models of all domains, and selected mathematics achievement as the only domain for treatment in the explanatory analyses in PISA 2003. The model indicated that, on average across the OECD, 46% of between-school variance was attributable to student and school SES (parental occupation, parental education, cultural home possessions, lone-parent status, country of birth, and language spoken at home). The OECD comments: "These findings have potentially important implications for policy-makers. ... countries in which the relationship between socio-economic background and student performance is strong do not fully capitalise the potential of students from disadvantaged backgrounds" (2004c, p. 174). The OECD also reported that school climate factors explained an additional 5% of between-school achievement variance, over and above student and school social background; school policies and admittance factors an additional 2%; and just over 1% by school resources. The conclusions drawn from these analyses are that schools can make a difference (albeit a relatively small one), and that social background and school characteristics covary.

What has *not* been considered in the multilevel models in the international reports on PISA, however, is the extent to which the sample design (age- or grade-based) and test content/area (whether aligned to the curriculum or not; whether school-dependent or

not) should be taken into account when interpreting the results. These issues are important, since, along with the between-school variance statistic, the OECD (2001b, 2004c, 2005c) takes the magnitude of the effects of social background (school social intake) as an indicator of the extent to which schools can moderate the impact of socioeconomic disadvantage. If the relative impact of social background were to vary substantially depending on the sample design and test content/area, then the conclusions one might draw about the extent to which schools moderate the impact of social background will also differ. Further, the extent to which school-level variables impact on achievement over and above social background may vary on the basis of both test content/subject area and sample design. If this is the case, then conclusions about the extent to which school-level variables affect student achievements will also differ depending both on the survey design and outcome measure used. I demonstrate in Chapter 2, through a review of published analyses of variance components and explanatory models of the achievements of Irish students, why both of these issues are of relevance and concern in the Irish context.

It should be noted that most multilevel models reviewed in Chapter 2 do not include an adjustment for student intake in the manner recommended by Goldstein et al. (1999). As a consequence, the effects associated with social background are likely to be overestimated, and, possibly, incorrectly ascribed to the school that the student currently attends (see also Nash, 2003). There are other general limitations to these models. First, in some studies, an attempt is made to quantify the 'value added' component of schools; i.e., the extent to which school-level or class-level variables explain variance in achievement after adjusting for social background of students and the social composition of schools. Raudenbush and Willms (1995) term these Type A and Type B effects, where Type A effects refer to the social background of students and the social composition of schools, and Type B effects refer to school or class resources or practices. They point out that adequate measures of Type B effects are more often than not absent from these models, due to poor measurement/conceptualisation (see also Smithers, 2004; Goldstein, 2000), or the limitations imposed by a cross-sectional survey (Type B effects might be better studied within a process model, i.e., effects across time examined within a longitudinal survey) (Goldstein, 2004). Goldstein and Spiegelhalter (1995) also note the limited availability of information on Type B effects and argue that "In the absence of good understandings about Type B effects, the distinction between

Type A and Type B effects is of little practical significance..." (p. 14). These reasons may help to explain why the school-level variables in models for PISA 2003 reviewed above explained so little achievement variance.

Second, it is difficult if impossible to make causal inferences. Here, the distinction between treatments and attributes is a useful one. Raudenbush and Willms (1995) draw on recent statistical theory regarding causal inference in their discussion of this issue (citing Holland, 1986; Rosenbaum & Rubin, 1983; Rubin, 1978). Essential elements in a causal study are two sets (i.e., a set of treatments to be evaluated, and a population assigned to these treatments) and two random variables (group assignment, outcome of treatment). A treatment must be capable of manipulation and its effect conceived in relation to alternative, available treatments. Therefore, student social background is an attribute, whereas a method of instruction is a treatment; Type A and Type B effects both arise from a complex mix of attributes and treatments. One can further split treatments into 'inputs' (e.g., pupil-teacher ratio; financial grants for library facilities) and 'processes' (e.g., amount of homework given; instructional methods) (Goldstein & Spiegelhalter, 1995). Raudenbush and Willms ask: "Given the impossibility of randomization in standard school evaluations, is unbiased inference possible?" (p. 312). Given that students are not randomly assigned to schools, causal inference is not possible, unless covariates related to the outcome that affect the propensity of a student to attend a given school are taken into account. Even then, Goldstein and Spiegelhalter recommend that "...we should exert caution when applying statistical models to make institutional comparisons, treating results as suggestive rather than definitive" (p. 24).

Third, in the case of multilevel models in international reports of cross-country comparisons of achievement, it is usually the case that the same variables are used in each country's model, regardless of whether the variables are relevant, reliable, and statistically significant (e.g., OECD, 2004c). Worse still, some multilevel models (e.g., OECD, 2001b) have pooled the results of all countries and produced a three-level model which examines variation between countries, schools and students. This not only suffers from the conceptual difficulty entailed in assuming that participating countries are a random selection from all possible countries, it provides little or no relevant information to individual countries.

Fourth, even within a country, there may be variations in the extent to which explanatory variables are valid or relevant. For example, Sofroniou, Archer and Weir (in preparation), point out that the appropriateness of indicators of social background can vary, depending on the urban/rural location of students/schools.

Given these limitations, multilevel modelling techniques are still potentially useful for describing the relationships between socioeconomic status and student outcomes. Willms (2002) uses the term 'social gradient' to describe this relationship. Socioeconomic status, or SES, has been defined as "the relative position of a family or individual in an hierarchical social structure, based on their access to, or control over, wealth, prestige, and power" (Willms, 2002, citing Mueller & Parcel, 1981). It is often operationalised as a composite of income, education and occupational prestige. Deaton (2002) argues that a combined measure of SES is not useful for policy, since there is no policy instrument that can act simultaneously on income, education and social class. However, Willms (2002) points out that composite measures of SES should be understood as proxies for the relative position of individuals in the social structure, and as such, are a useful device for communicating the nature and extent of social inequalities in a society. He recommends against including variables in an SES composite which are not part of the formal definition of SES, such as family structure or family size. He also recommends examining gradients first by using a composite measure, and then examining the relationship between outcomes and constituent components as well as other factors such as ethnicity and family structure.

Turmo (2001) has discussed the conceptualisation and measurement of SES status in PISA, in which its questionnaire framework distinguishes several components of SES: occupational status [on a measure which combines education and income (the International Socio-Economic Index; see Ganzeboom & Treiman, 1996)]; cultural capital (based on Bourdieu's 1973, 1984 theory of cultural reproduction); and social capital (based on the work of Coleman, 1988). The SES variable which has been most widely used in the OECD reports of PISA is a composite measure of economic, social and cultural status (ESCS), which combines parental education, parental occupation, and educational and material resources in the home (e.g., OECD, 2001b; 2004c). Hence, the notion of 'social capital' as described by Coleman (1988), namely the resources provided through social ties, is absent from the manner in which the PISA measures

have actually been operationalised, which would appear to combine aspects of economic, occupational and cultural status or educational climate only. Whether this is the optimal operationalisation of SES or not is unclear. What is also unclear is whether the ESCS measure is appropriate and valid in all participating countries.

Willms (2002) notes that the strength of the social gradient and its functional form (whether linear or curvilinear) can vary with the unit of analysis (e.g., whether individual, school or community) and argues that "much more can be learned about the underlying processes that affect social outcomes [such as literacy skills] through a careful examination of gradients at each level of analysis" (p. 2). Given the limitations of the models regarding causal inferences, however, one should view such models as merely broad indications of the extent of risk associated with varying social backgrounds and where students at greatest risk are likely to be.

In sum, there are considerable limitations to the interpretations that can be made from multilevel models that attempt to explain student outcomes, particularly in the absence of adequate adjustments to take into account the prior characteristics of students. While such adjustments should be made both on the basis of general scholastic ability and social background, as a minimum, the technique is still useful for examining differential outcomes associated with SES, and how SES operates at individual and group levels. In the absence of adequate adjustments though, such models are best viewed as descriptive and diagnostic, indicating social inequalities that have a history beyond that of the student's current context, rather than explanatory and relating only, or mainly, to the student's current context. In addition, in the case of the PISA models reported by the OECD, no account has been taken of the potential impact of the sample design and test measure on the results, and this rather 'woolly' approach makes policy development on the basis of the results difficult.

1.8. Importance of PISA in Ireland

The importance of PISA in educational policy and public debate on the education system is evident when one considers the role of PISA in Ireland's system for monitoring educational achievements and its relative prominence in media commentary. Despite this, however, no critical commentary of the PISA survey as it relates to the Irish education system has been published, other than commentary on PISA as it relates

to the mathematics curriculum (e.g., Close & Oldham, 2005; Oldham, 2002). In contrast, there has been some critical analysis of PISA in the UK (e.g., Goldstein, 2004; Nash, 2003; Prais, 2003; Smithers, 2004). The lack of a general critical commentary in Ireland is somewhat surprising given some of the issues raised earlier in this chapter. This section describes PISA's role in the system for monitoring educational achievements in Ireland and reviews media and government commentary on the results.

1.8.1. The Role of PISA in Ireland's National System for Monitoring Educational Achievements

Kellaghan (1995) has developed a system for monitoring educational achievements for Ireland in a policy climate which, similar to developments internationally, emphasised objectives, standards, targets and accountability of the education system (Kellaghan & Greaney, 2001; Kellaghan & Madaus, 2000).

The monitoring system in place in Ireland prior to this was unsatisfactory because assessments were not systematic or regular, and there was a lack of trend data in many subject areas. Taking these issues into account, Kellaghan (1995) developed a system of monitoring educational achievements which fulfils criteria identified by Greaney and Kellaghan (1996). It:

- incorporates assessment at both primary and post-primary level;
- does not overburden the education system;
- ensures Ireland's continuing participation in PISA whilst also incorporating nationally tailored assessments;
- is capable of describing the output of an education system at a particular point in the system;
- is capable of identifying areas of knowledge/skills which are deficient, suggesting problems in curriculum implementation;
- is capable of examining achievement by gender, location and other policy relevant contextual variables; and
- is capable of producing trends, i.e. monitoring changes in achievement over time.

The schedule of data collection for this monitoring system is outlined in Table 1.6. It is evident that the nature of monitoring differs at primary and post-primary levels. At

primary level, national assessments have been designed to measure aspects of the intended curriculum (see Eivers, Shiel, Perkins, & Cosgrove, in preparation; Harris, Forde, Archer, Nic Fhearaile, & O'Gorman, in preparation; Shiel & Surgenor, in preparation). PISA is at present the only assessment which monitors the education system at post-primary level and it hardly need be mentioned that the PISA tests are not tailored to the national curricula. A second difference is that assessments at primary level are anchored at a particular class level or point in the system, and the surveys use intact-class sampling. The PISA sample is not anchored to a particular point in the system, although the modal grade of Irish students participating in PISA is, conveniently, third year, i.e., the end of Junior Cycle. A third difference is the asymmetry in the monitoring system. At post-primary level, conclusions may be drawn about Ireland's performance with respect to international standards, but not with respect to national ones. The reverse is true at primary level. This asymmetry, together with PISA's curriculum-free approach to assessment, point to the potential relevance of further analyses at national level to provide a better understanding of what the PISA standards mean in Ireland.

Table 1.6. Educational Assessments in Ireland's Monitoring System, 1998-2009

National/International?	*Age/Grade*	*Subject Area*	*Year*
National	11 years/Fifth Class	English Reading	1998, 2003*, 2008
National	10 years/Fourth Class	Mathematics	1999, 2004, 2009
National	12 years/Sixth Class	Irish: Aural, Oral and Reading	2002, 2007
International	15 years/Third Year	Reading Literacy (major focus)	2000, 2003, 2006
International	15 years/Third Year	Mathematics (major focus)	2000, 2003, 2006
International	15 years/Third Year	Science (major focus)	2000, 2003, 2006
National	?/? (post-primary)	Irish	Unknown at present

*Due to the requirement to develop a new reading test for first class pupils and update the 1998 fifth class test, the 2003 assessment took place in 2004. First class as well as fifth class form the target population.
Source: Kellaghan (1995), Table 1.

1.8.1. Media and Government Commentary on PISA

The significance of PISA in Ireland becomes even more evident when one considers media reports and commentary from politicians. For example, the PISA survey is described in national newspapers as a "major new international survey" (Oliver,

December 5, 2001) and "the biggest ever international study of student achievement" Walshe, December 5, 2001); and the OECD as a "prestigious international body" (Oliver, December 5, 2001) and "the Paris-based think-tank" (Walshe & Donnelly, September 17, 2003).

Emmet Oliver states that

> The report [*Education at a Glance*] is the main source of information on education standards and performance in the industrialised world. Governments throughout Europe and elsewhere take its findings extremely seriously… . (October 30, 2002)

In an *Irish Times Editorial*, it is stated that

> Teachers, parents and the Government will be closely examining the findings of this year's OECD report on educational standards in Ireland and throughout the industrialised world. (October 30, 2002)

Despite the apparent importance of the survey, however, the vast majority of the articles simply report the results in a factual manner, using the results to criticise or praise the education system rather than questioning the nature of the survey. In almost every newspaper article on PISA, the emphasis is on country rankings and/or Irish performance with respect to the OECD average, e.g.:

> Irish students rank second highest for reading ability in Europe and fifth highest in the world, an analysis of international education trends has revealed. (*Irish Times Editorial*, December 5, 2001)

Only one newspaper article has mentioned the measurement error associated with the country rankings:

> The authors of the [PISA 2000] report state that care should be exercised in interpreting outcomes and, that a country's rank is "a crude measure" of performance. They go on to state that "interpreted with care, such information can provide valuable insights into a country's education system in a comparative context". (*Irish Times Editorial*, January 15, 2002)

The percentages of students at the upper and lower ends of the proficiency scales are also often cited. However, there is evidence that the distinction between Level 1 and

below Level 1 is being interpreted in different ways, as the following two excerpts demonstrate. This is potentially quite a serious error since there is a substantial difference between the percentage of students at *and* below Level 1, and those below Level 1 only:

According to the Organisation for Economic Co-operation and Development (OECD) survey, less than 3 per cent of school-goers here showed serious literacy problems [i.e., are below Level 1], a significantly better result than most of the developed economies surveyed. (*Irish Times Editorial*, December 5, 2001)

Just 11 per cent of Irish students achieved scores in the lowest category [i.e. at and below Level 1], compared with an OECD average of 18 per cent. (Healy, September 17, 2003)

Many of the articles make the link between achievement and economic competitiveness, e.g.:

...over 17 per cent of 15-year-olds are scoring at the lowest possible proficiency level in maths, indicating that they have insufficient skills to meet their own future needs and the needs of society. The Government and business believe higher maths standards are necessary if the Republic is to realise its ambition as a leading Knowledge Society. (Flynn, December 7, 2004)

In a disturbing finding, the report states that more than 17 per cent of Irish students are scoring at the lowest proficiency level [in PISA mathematics]. This indicates, according to the report, that they have insufficient skills to meet their own future needs, let alone those of society. High achievement in the subject is seen as an essential building block as the Republic seeks to progress towards the much-vaunted "knowledge society". But the latest OECD research highlights the scale of the challenge. (*Irish Times Editorial*, December 7, 2004)

In a group of eight countries that achieved significantly higher-than-average scores in reading literacy, Ireland was alone in not receiving a similarly high score in maths. Low attainment in this area could have serious consequences for a country's competitiveness and labour market earnings... . (Healy, September 17, 2003)

Between-school variance has also been mentioned and interpreted to mean that the Irish system is relatively high in equity, e.g.:

Despite the perception that the Irish system is rife with inequality, the OECD report says the difference in performance among schools is not large compared to other countries. (Oliver, December 5, 2001)

There has been no media commentary with respect to the English curriculum, which is viewed as unproblematic; due, probably, to the high average performance of students; however there have been calls for curricular reform in mathematics: "Given the strong evidence presented [in the PISA 2003 mathematics results] policymakers may have no choice but to review the maths syllabus once again" (*Irish Times Editorial*, December 7, 2004).

Government ministers in Ireland have also taken note of the results:

> Studies such as PISA provide important information to enable the performance of Irish students and the Irish education system to be benchmarked against international trends. I am very pleased that Irish 15 year olds were among the top performers with regard to reading in PISA 2003 as in the 2000 study and that they have maintained their position in mathematics and science. The results of PISA 2003 give us much to celebrate with regard to the achievements of our education system and they also highlight challenges which we will need to work towards addressing. I look forward to the more detailed analysis which will emerge in the coming months as the PISA results are examined in greater depth. (Minister Hanafin, quoted in a Department of Education and Science press release, December 7, 2004)

Indeed, the percentage of students at or below Level 1 is included as a key indicator in the recently-published *Key Education Statistics – 1993/94-2003/04* (Department of Education and Science, 2005). In a press release from the Department of Education and Science announcing the publication of *Key Education Statistics* (August 25, 2005) it is noted that

> ...19.8% of 15-year-old pupils in the EU countries participating in the OECD PISA survey were found to be low achievers in reading literacy compared to the EU Benchmark for 2010 of 15.5%. Ireland was the 2nd best performing EU country with only 11 percent of 15 year olds categorised as low achievers.

The Minister for Education and Science at the time of PISA 2000 is quoted in a Department of Education and Science press release (December 7, 2001) as saying:

> I am particularly pleased that Ireland achieved 5th place in reading literacy and 9th place in scientific literacy of the 28 OECD countries that participated and that our scores in these areas were significantly higher than the average OECD score.

The press release mentions the percentages of Irish students at or below Level 1 on reading in PISA 2000: "fewer Irish students achieved scores at the lowest level of proficiency (11%) in reading than the OECD average (17%)". The press release also states that "the emphasis on the assessment of students' preparedness for life in terms of the knowledge and skills required for their future lives is of particular significance".

Minister Woods also took note of the OECD's analyses of PISA 2000 relating to socioeconomic status and it is stated that

> ...the outcome of the analysis of various socioeconomic variables at school and student level and their effects on student achievement [is noted]. He [Woods] said that these influences on student performances will be analysed in his Department in the context of policy development for educationally disadvantaged students.

The dual themes of quality and equity with regard to socioeconomic status have also been emphasised at the level of OECD ministers. In the Chair's summary of the meeting of OECD education ministers (Dempsey, 18-19 March, 2004), it is stated that countries were in agreement that

> ...PISA results, by showing that some countries are successfully combining high performance standards with a socially equitable distribution of learning opportunities, had sent an important and encouraging message for all countries, namely that poor performance does not automatically follow from social disadvantage.

1.9. Conclusion

The political agenda underlying surveys are important determinants of their aims. The agenda underlying PISA are quite different to previous surveys of the IEA, since the focus is on the economic success and competitiveness of market economies rather than the quest for theoretical and methodological progress in comparative educational research.

The current interest in educational standards has perhaps never been higher. PISA represents a significant progression in the work of INES and in the arena of comparative educational research since it has resulted in the availability of achievement data and other educational indicators for all OECD countries (plus an increasing number of 'partner' countries) which have been collected using rigorous technical standards,

comparable to, or even surpassing, those of previous international assessments (Goldstein, 2004; International Reading Association, 2003; Smithers, 2004).

However, there are problems in the manner in which PISA's survey aims and objectives are translated into its design, and these have significant implications for the interpretation of results. The problems centre on the often-debated areas of test content and sample design. Given PISA's departure from the previous tradition of attempting to measure aspects of participating countries' curricula, however, a new slant is put on the problems.

First, there is, somewhat of a mismatch between some of PISA's objectives and the manner in which student achievements are assessed (see DfES, 2004; Goldstein, 2004; Prais; 2003; Smithers, 2004). The mismatch pertains to the OECD's claim that results of PISA can be used to establish benchmarks for educational improvement, and to understand relative strengths and weaknesses of educational systems. Taking a literacy-based approach, and basing the tests on what knowledge and skills are desirable for young adults side-steps rather than solves the issue of curriculum variation across participating countries, which is likely to be more of an issue in the interpretation of mathematics achievement than reading achievement. This has been expressed by Smithers as a 'curious contradiction', whereby countries need somehow to interpret the results to make improvements to their education systems in the absence of comparative information about curricula and how these relate to PISA. About eight countries have taken this issue up with analyses of the PISA tests with respect to their national curricula, but the results are not comparable and the OECD Secretariat and PISA Governing Board appear to be resistant with regard to supporting the development of an international comparative framework in which to interpret curricula (DfES, 2004).

Second, criticism has been levelled at claims by the OECD that students' performance on a literacy test is a valid measure of the relevant skills needed by market economies, and questions have been raised about the predictive validity of studies such as PISA and the appropriateness of generalising, on the basis of results of a literacy test given to students, to the relative economic strength and competitiveness of individuals (and the entire country) (Bonnet, 2002; Kellaghan, 1996). However, a re-analysis of IALS data (which itself suffers from methodological shortcomings) suggests, at the country level,

that gains in literacy are associated with gains in economic performance (Coulombe et al., 2004). The re-analysis found that a reduction in the percentage with low literacy levels was particularly strongly associated with economic gain. However, there are still no supporting data on predictive validity at the individual level and no obvious means, within a cross-sectional design, of supporting this claim.

Third, while the arbitrary nature of the cutpoints established with proficiency levels is stated in the technical documentation, it is not in the main OECD reports on PISA. This is important since the percentages of students at various proficiency levels has received a relatively high focus in the media and by government ministers in Ireland and value judgements and absolutist statements regarding benchmarks and standards have been attached to these results. Furthermore, although the basis of establishing cutpoints for PISA is grounded in defensible principles, and detailed descriptions accompany each proficiency level, the combining of Level 1 with below Level 1 may not be justified and may disguise between-country (and within-country) differences in the proportions who are already at Level 1 with the proportions below Level 1. Also, due to the low numbers of items at the extremes of the ability distribution, PISA is rather limited in its description of the knowledge and skills students achieve at the extremes. If the conclusions of Coulombe et al. (2004) are valid, then increased precision at the lower end of the literacy scale to allow enhanced monitoring of the capabilities of lower achievers would be desirable.

Fourth, while an age-based sample as used in PISA addresses the issue of what knowledge and skills students in OECD countries nearing the end of compulsory schooling have acquired, it was noted that the compulsory schooling age in the majority of OECD countries is actually 16, not 15. Also, the samples still vary considerably according to grade dispersion, whether in a lower secondary or upper secondary programme, etc. Furthermore, there is a lack of international data on grade repetition. Therefore, while PISA students may be near the end of compulsory schooling, these differences make it extremely difficult in most countries to pin achievements to a particular point in the system and compare their results with confidence with those of other countries.

Fifth, the between-school variance statistic is often taken as a measure of the homogeneity of schools in terms of achievement outcomes (Postlethwaite, 1995). The OECD has taken this interpretation further in the reports on PISA, citing the statistic as an indicator of educational equity. It was suggested, however, that the sample design (whether age- or grade-based) can impact significantly on this statistic, particularly with respect to the mechanisms used to allocate students to schools and classes, and should be considered in its interpretation; a detail not present in the OECD reports on PISA. A further detail not considered in the reports is the possibility that between-school variance may be related to whether the achievement measure is aligned to the curriculum and whether the subject domain is school-dependent or not.

Sixth, more recently, concerns have been expressed about the extent to which non-response and school and student exclusions bias the survey results. O'Leary et al. (2001) cite this as a 'vexing' aspect in the interpretation of results, and other writers (e.g., Beaton et al., 1999) recommend that this bias needs to be quantifiable. By minimising non-response due to exclusions (with strict standards for population coverage, school-level and student-level exclusions) and the use of a one-hour test booklet for countries with more than 5% of 15-year-olds enrolled in schools for students with special needs, the bias relating to exclusions is kept to a minimum in PISA. However, some analyses using simulated datasets have demonstrated the existence of a bias arising from non-response which affects both estimates of average achievement and of achievement variance. The extent of bias appears to vary according to how achievement variance is partitioned between and within schools (Monseur & Wu, 2002). There are calls for standards on the quantification of bias (DfES, 2005).

Seventh, it was pointed out that increasingly, multilevel explanatory models of achievement are used in reports of international surveys. These have the advantage over previously-used regression techniques in that they account for the nested nature of the sample design. However, there are general problems with drawing inferences from these models relating to (i) the absence, in many, of an appropriate adjustment for intake, (ii) difficulties in distinguishing clearly between effects relating to social background and other effects relating to school practice, and/or the lack of availability of strong measures of school practice, and (iii) a lack of comparability in the conceptualisation and measurement of social background or socioeconomic status. The importance

ascribed to socioeconomic status and the social context effect in recent international survey reports was noted. Analyses along these lines have also investigated the extent to which, once achievement is adjusted for by social intake of schools and social background of students, school/class variables explain achievement. However, these reports do not consider the extent to which the content of the test, the subject area, or sample design impact on the apparent associations between socioeconomic and school/class factors and achievement variance. Moreover, many of the explanatory variables used suffer from weak conceptual underpinnings and/or poor explanatory power.

Finally, the importance of PISA in Ireland is evident, both in terms of its role in monitoring the education system at post-primary level, and the attention the results have received in the media and by government ministers. Public commentary, however, has been uncritical of the approaches and design underpinning PISA. The focus, rather, tends to be rather simple, based on country rankings, the percentages achieving at the lowest proficiency levels (and some confusion with the interpretation of these was noted), the possible impact of the results on the economy, and the impact of socioeconomic factors on achievement.

CHAPTER 2. A CONSIDERATION OF PISA IN THE IRISH CONTEXT

2.1. Introduction

This chapter considers PISA's achievement results in the Irish context with respect to four aspects identified in Chapter 1 as worthy of more in-depth investigation.

First, given the importance placed by the media and government on the percentages of students at and below a minimum level on the PISA proficiency scales, and tentative evidence from re-analysis of the IALS data that the monitoring of the lower end of the achievement tail in particular may be of importance as an indicator of a country's economic growth potential, the precision of such estimates is important. However, there is some evidence of bias in achievement estimates arising from non-response which is not controlled for, even if achieved samples attain specified quality standards. Further, the efficiency of statistical adjustments varies across countries according to the percentage of variance between/within schools. Therefore, the first section considers the potential for bias in estimates of mean achievement and variance in achievement arising from non-response with respect to the Irish PISA datasets. This issue will be explored further in analyses in Chapter 3.

Second, given that PISA departs from previous surveys in its move away from attempts to assess achievement with reference to curricular content, and the fact that the PISA achievement measures are the only source at post-primary level of data that allow international comparisons of educational outcomes, the results of existing analyses which compare the content of, and performance on, PISA and the Junior Certificate, are reviewed (and some limitations of these analyses are noted). The aim is to reach a judgement as to what the PISA measures can tell us about achievement in Ireland, to identify further analyses which might add to an understanding of how achievement on PISA relates to performance on national examinations (the Junior Certificate), and to consider whether the design of PISA itself might need to be modified to enhance the interpretability of results in this regard. Results of additional analyses of links between PISA and the Junior Certificate are described in Chapter 4.

Third, on the theme of the equity of achievement outcomes, a review of published analyses of between-school/class variance in the achievements of Irish students is

undertaken in order to identify patterns in the manner in which achievement variance is partitioned between schools which may be linked to the sample design and/or the nature of the achievement measure. The review is used to develop a number of research questions which are explored in Chapter 5 through a comparison of variance components of PISA, TIMSS and the Junior Certificate.

Fourth, in considering the determinants of achievement, explanatory models of achievement in Ireland are reviewed in order to identify common patterns in the results. Limitations are noted regarding the ability of the existing explanatory analyses to address the issues raised about the impacts of social intake and school/class practice variables. This information is used to develop a number of hypotheses as to how both the nature of the test used and the sample design should be considered when interpreting results of explanatory analyses of social background and school/class effects on achievement. Chapter 5 documents the results of explanatory models of PISA, TIMSS and the Junior Certificate which attempt, using the best available data, to address the questions raised.

This chapter concludes by identifying three core questions are identified and describes how the analyses in Chapters 3, 4 and 5 address these.

2.2. Evidence for Bias Arising From Non-Response in Ireland

2.2.1. Evaluation of the Achieved Sample for Ireland in PISA 2000 and PISA 2003

Before discussing the potential for non-response bias in the Irish samples for PISA 2000 and PISA 2003, I want to demonstrate that the Irish sample met all required sampling standards as described in Chapter 1 (Section 1.7).

In PISA 2000, Ireland's population coverage index was .95 (Monseur, Rust, & Krawchuk, 2002, Table 31, Index 3). In PISA 2003, population coverage for Ireland was similar, at .96 (OECD, 2004c, Table A3.1, Column 15). These indicate that Ireland's coverage of the population met the 95% criterion. Within-school exclusions of students with special educational needs were also at an acceptable level (3.1% in 2000 and 2.7% in 2003) (Monseur, Rust & Krawchuk, 2002, p. 135; OECD, 2005b, pp. 168-169).

Some replacement schools participated in both surveys (three in PISA 2000 and four in PISA 2003). Nonetheless, in both years, Ireland exceeded both the school-level and student-level response rate standards, with a weighted response rate of 85.6% before replacement and 87.5% after replacement at the school level, and a weighted response rate of 85.6% after replacement at the student level in PISA 2000. The weighted response rates for PISA 2003 are also acceptable (90.2% school-level response rate before replacement; 92.8% school-level response rate after replacement; within-school response rate of 82.6%).

2.2.2. Non-Response and Mean Achievement

There is ample evidence to show that student non-response is not random in many of the countries participating in PISA 2000, including Ireland, even though the weighting process assumes that it is. Monseur and Wu (2003) obtained a Pearson correlation of .32 between the average (aggregated) school achievement in PISA 2000 reading literacy and the percent of (eligible) students participating in each school in Ireland. Replicating this analysis using an unweighted dataset in SPSS, I obtained a correlation of .33. This was computed by assigning each eligible student a value of 0 or 1 depending on their participation status and aggregating this to the school level, resulting in a proportion ranging from 0 to 1, which corresponds to eligible student response rate per school. I then averaged the five reading literacy plausible values and aggregated these to the level of the school. The Pearson correlation between these two aggregated variables yields the measure of association between propensity to participate and school reading achievement. A correlation of .32 or .33 is quite high relative to the other participating countries in PISA 2000 (6th highest out of 31 countries), where correlations ranged between -.02 to .53, with a country average of .19.

Using the same technique with the PISA 2003 dataset, but this time averaging the mathematics scores, I obtained an even higher correlation of .40. These correlations confirm that the response rates *within* schools in Ireland, in general, are positively associated with the average proficiency of students in schools in both PISA 2000 and PISA 2003.

Monseur and Wu (2003) also reported the outcomes of analyses of simulated datasets on the effects of student and school non-response adjustments as used in PISA on the

extent of bias in achievement estimates, using an expected student response rate of 80%. They found that the efficiency of both the school-level and student-level non-response adjustment was proportional to the between-school variance in achievement. The higher the between-school variance, the more efficient the student-level non-response adjustment; the lower the between-school variance, the more efficient the school-level non-response adjustment. In Ireland, the between-school variance was comparatively low in both 2000 and 2003. This suggests that the student-level non-response adjustment is likely to be inefficient, and may result in a bias in achievement estimates. This, coupled with the fact that the correlation at the school level between participation rates and achievement is comparatively high, suggests that the student-level estimates for Ireland may be significantly upwardly biased. In contrast, it suggests that non-response adjustments at the school level for Ireland are quite efficient.

Differential student participation rates were observed by Monseur and Wu (2003) by gender, age, and grade level in many of the countries participating in PISA 2000. This becomes problematic if any of these variables is related to achievement *and* if there is differential participation associated with the variable. In Ireland, grade level is related *both* to achievement (see Shiel et al., 2001, Table 4.27; Cosgrove et al., 2005, Table 4.19) and differential participation, according to Monseur and Wu's analyses of the PISA 2000 data. Monseur and Wu did not find an association between gender and participation, nor one between age and participation, in Ireland.

Other variables not analysed by Monseur and Wu (2003) but nonetheless relevant to the interpretation of achievement estimates in Ireland with respect to differential school response rates include the two implicit stratifying variables[13]. Performance on PISA 2000 reading differed significantly across schools of differing sex composition, with students in all girls' schools (M = 549, SE = 5.7) achieving a higher mean than students in boys' schools (M = 533, SE = 6.1) and in co-ed schools (M = 516, SE = 4.6). In PISA 2003, mathematics achievement was higher in boys' schools (M = 529, SE = 5.6) than in girls' schools (M = 507, SE = 5.6) and also higher than the two categories of mixed sex schools (low female mixed M = 481, SE = 6.1; high female mixed M = 499, SE = 4.3).

[13] Means and standard errors cited in this section are taken from Shiel et al., 2001, Chapter 4 for PISA 2000; and from Cosgrove et al., 2005, Chapter 4 for PISA 2003.

Achievements on PISA also differ with respect to school sector. On the PISA 2000 measure of reading literacy, students in secondary schools (M = 543, SE = 3.8) achieved significantly higher reading literacy scores than those in community/comprehensive schools (M = 522, SE = 6.4) and than those in vocational schools (M = 484, SE = 6.7). Similarly, students in secondary schools in PISA 2003 (M = 514, SE = 3.3) scored significantly higher on the mathematics test than students in both community/comprehensive schools (M = 498, SE = 5.1) and vocational schools (M = 474, SE = 5.5).

Student performance on PISA 2000 reading also varied significantly across designated disadvantaged (M = 490, SE = 7.4) and non-designated schools (M = 539, SE = 3.1). The performance difference is also significant for mathematics in PISA 2003 (designated schools M = 477, SE = 4.8; non-designated M = 512, SE = 2.8).

A student-level variable in addition to gender and grade level which is particularly pertinent to the present research is the syllabus level at which the student took Junior Certificate English (in the case of PISA 2000 analyses) and Junior Certificate mathematics (in the case of PISA 2003). In Section 2.3, it is shown that the average performance of students on PISA varies substantially across syllabus levels in both English and mathematics.

Thus, if student participation rates were to differ significantly across these variables, a bias in the achievement estimates that is not corrected by the non-response adjustments is to be expected. Given that some of the achievement differences relate to school-level variables, a comparison of non-response at the school is also warranted, even though Monseur and Wu's (2002) analyses suggest that school non-response in Ireland is unlikely to give rise to bias.

2.2.3. Non-Response and Variance in Achievement

A second consequence of student non-response suggested by Monseur and Wu (2003) is a bias in estimates of the variance in achievement, which may manifest itself in two ways. First, if student propensity to participate is positively related to achievement, then *schools* may appear more homogenous than they actually are, had all eligible students

participated. It was already noted that student propensity to participate is significantly related to achievement in Ireland, and among the highest of the countries participating in PISA 2000. Further, in both PISA 2000 and PISA 2003, the between-school variance in achievement in Ireland was comparatively low. It is reasonable to hypothesise that full student participation rates might have resulted in higher between-school variance.

Second, variance between *students* is likely to be reduced if proportionately more lower achievers did not participate, since the between-student variance of participating students may not incorporate some of the individual student variance at the lower end of the achievement distribution, had all eligible students participated. This is pertinent in the case of Ireland since Irish achievement on PISA is characterised by comparatively low between-student variance, as indicated by the standard deviation. In PISA 2000, the standard deviation for reading literacy for Ireland was 93.6, which is below the OECD average of 100.0. In PISA 2003, the standard deviation for mathematics for Ireland was 85.3, again below the OECD average of 100.0. Moreover, in both PISA 2000 and PISA 2003, the within-school response rates for Ireland, although above the specified minimum of 80%, were lower than average. In 2000, the weighted within-school response rate for Ireland was 85.6%, which is lower than the average of the 31 participating countries (90.6%). In 2003, the weighted within-school response rate for Ireland was 82.6%, again lower than the average (90.9%) of the 40 participating countries.[14]

2.3. Existing Research on PISA and Curriculum in Ireland

Existing research on PISA and the curriculum in Ireland comprise broad, qualitative comparisons of the Junior Certificate syllabuses and examinations and the PISA assessment frameworks and tests; a quantitative analysis of PISA test items with respect to the Junior Certificate; and a comparison of actual performance on PISA and the Junior Certificate. Later, it will be shown how the analyses in Chapters 3 extend the existing research. It should be noted that, while many of the analyses reported in this section draw attention to differences in the content of the syllabuses at higher, ordinary and foundation levels, and while achievement differences across the syllabus levels on PISA are substantial, little is known about how students come to select (or are selected

[14] The country average within-school response rate and OECD average response rate were calculated as the arithmetic means.

for) the examinations at various syllabus levels. Further, a review of the syllabus documents and teacher guidelines for both subject areas[15] suggests that there is a lack of concrete guidelines for schools and teachers as to which types of students might by suited by the various syllabus levels.

2.3.1. Qualitative Comparisons

The national reports for PISA 2000 and PISA 2003 (Cosgrove et al., 2005; Shiel et al., 2001) included a description of the Junior Certificate syllabus and examinations for English, mathematics and science; the descriptions for English (2000) and mathematics (2003) only are reviewed here.

2.3.1.1. Junior Certificate English[16]

In the Junior Certificate English syllabus in place at the time of PISA 2000 (Department of Education, n.d.), a distinction is made between personal, social, and cultural literacy, which is consistent with the multiple functions and contexts of reading evident in PISA (OECD, 2000b). Understanding of and expression through aesthetic texts, as well as understanding, using and producing public, functional texts are mentioned in the syllabus. Texts are distinguished the basis on their intended purposes and audiences. The use of a diversity of text types in instruction is mentioned a number of times. Teachers are encouraged to select texts which are felt to be appropriate to the cultural environment, stage of development, and linguistic abilities of their students. A holistic and integrated approach is emphasised. Teacher guidelines specify a list of targets and activities for each of the three years of the Junior Cycle for the syllabus strands of language, literature, oral, aural, reading, and writing skills (Table 2.1).

[15] Syllabus documentation is available at http://www.ncca.ie

[16] This review of the Junior Certificate English syllabus and examinations, with the exception of the commentary on the marking schemes, is based on the review of Shiel et al. (2001).

Table 2.1. Junior Certificate English (1989 Syllabus): Strands, Targets, and Activities, by Year Level

Strand	*First Year*	*Second Year*	*Third Year*
Language	Develop an understanding of basic forms and structures of sentences and paragraphs; develop basic punctuation conventions; have lexical awareness, have a sense of audience.	Develop an understanding of: forms and structures of longer compositions; basic punctuation conventions; more complex spellings; more challenging sense of audience and purpose; and lexical awareness.	Develop an understanding of vocabulary to discuss language use (e.g., connotation, cliché); manipulative language techniques; appropriateness of style and register; strategies for spelling and punctuation.
Literature	Understand and use the following: hero/villain, conflict, tension, climax, point of view, characters and relationships, scenes and story-shape, sound, texture and rhythm, style and word selection, and sensationalism/realism.	Understand and use the following: contrast, narrative voice, character development and motivation, beginning/end; mood, atmosphere, tone; style, word-pattern, and verbal choice. Literary forms of short story, novel and play.	Understand and use the following: plot, comedy, tragedy, satire, pathos, melodrama, theatre, lyrical, narrative, tone, irony and symbolism.
Oral/Aural	Encouragement to: tell an anecdote; have small group discussion; describe and report on events, places, people; interview and question; comment on television or radio programmes, and simple dramatic improvisation.	Encouragement to: record and dramatise narrative; engage in debates; give short speeches; ask questions in public lecture settings; discuss and evaluate media experiences, and present short dramatic scenes from texts.	Encouragement to: talk and listen in a wide range of contexts, both formal and informal, building on the activities of the previous two years.
Reading	Encouragement to: read own and others' written work for revision and editing purposes; read silently for a variety of purposes; use reference resources; read newspapers, and watch television programmes; attend to word choice, images and presentation; read a variety of literary genres with an awareness of sound, texture and rhythm.	Encouragement to: read silently for a more sustained period; engage in independent reading; read newspapers, journals, attending to viewpoint, assumptions, accuracy and style; contrast and evaluate different print media; comment on use of illustrations; view TV programmes and comment on implicit values; and read widely.	Encouragement to: identify types of order (e.g., chronological, spatial, importance); identify a writer's purpose; draw conclusions, predict outcomes, and suggest implications; be aware of narrative stance of the writer; distinguish between fact and opinion, and identify material which contains the language of stereotyping.
Writing	Procedures emphasised: prewriting, writing, rewriting and editing. Encouragement to: give information in note form; compose captions and titles; fill in application forms; report on an event; write personal and business letters; keep a diary; write simple dialogue and verse; and review literature, films and television programmes.	Encouragement to: develop the craft of writing; write reports; write formal letters; devise application forms, advertisements and brochures; write descriptive and argumentative essays; compose alternative scenes in literary texts; write in various literary forms; and evaluate a range of literary and media experiences.	Encouragement to: write more extended compositions in a wide range of contexts; and show a clear awareness of audience, purpose and register.

Source: Shiel et al., 2001, Table 6.1.

Although these strands are closely interlinked, it is reasonable to say that PISA does not assess oral/aural skills, and little writing skills. The targets and activities indicate that, by the end of the Junior Cycle, students of average or above average ability should have a well-developed set of skills and techniques for the critical reading, writing, and analysis of the structure, form, style, and tone of a wide variety of text types.

Formal assessment of the Junior Certificate English syllabus is in the form of written examination at three levels: higher, ordinary, and foundation. Shiel et al. (2001) have noted that the teacher guidelines differentiate only in very broad terms between the targets and activities expected at higher and ordinary levels, while guidelines for foundation level have not been published. Students respond to both unseen and studied material in the Junior Certificate English examination. Coursework includes studying a prescribed set of poems, short stories, plays, and novels. Both modern and classic texts are included. Students are assessed in the following areas (Department of Education and Science, n.d.; Shiel et al., 2001):

- Understanding and conveying information;
- Understanding facts, ideas and opinions;
- Analysing, evaluating and selecting relevant information for a given purpose;
- Describing and reflecting on experience (fictional and non-fictional);
- Recognising explicit meanings and some simpler implicit meanings;
- Expressing responses to a variety of literary genres;
- Showing a sense of audience; and
- Using appropriate spelling and punctuation.

Students at all levels are taught the same broad processes or skills, but differ in the depth and type of coverage, as well as the length, density, and complexity of the texts studied.

The content and structure of Junior Certificate Examination papers provide further information about the types of tasks that students are expected to do. The tasks encountered by students at each level in the 1999 examinations are described in Shiel et al. (2001) and summarised in Table 2.2. These suggest that the responses required are generally longer than the PISA tests and the emphasis on functional texts is lower.

Table 2.2. Description of Content of the 1999 Junior Certificate English Examination Papers, by Syllabus Level, Text, and Task

Foundation	*Ordinary*	*Higher*
	Section 1: Reading	
Text: Four short paragraphs about spiders. Expository.	Text: Five short paragraphs about snakes. Expository.	Text: A one-and-a-half page extract from a Bill Bryson novel (travel writer), containing a lot of southern US slang and dialect.
Tasks: Two questions requiring retrieval of information, two questions requiring inference regarding word meaning, one question requiring judgment regarding suitability of title.	Tasks: One question requiring retrieval of information, one question requiring interpretation and inference, one question requiring students to infer reasons for word choice, one question requiring students to comment on the writing devices used to convey mood, and one question requiring students to infer something about the author.	Tasks: One question requiring inference about the author as a person, one question requiring inference of attitude and feelings of the characters and one question requiring students to identify and comment on humorous devices in the text.
	Section 2: Personal Writing	
Text: Seven possible composition titles (e.g., *When I Was Small*).	Text: Seven possible composition titles (e.g., *My First Job*) and a line drawing.	Text: Eight descriptions of possible compositions (e.g., Imagine you are present at a great event in history. Write out in diary form your personal reactions to the event). Students are free to write in any form (e.g. dramatic, short story etc.).
Task: Write a page on one of the topics.	Task: Write a composition (length unspecified) on one of the topics (titles or drawing).	Task: Write a composition (length unspecified) on one of the topics.
	Section 3: Functional Writing	
Text/Tasks: One of A or B. A: Requirement to give a talk to pupils in 6th class about five problems they will have when entering post primary school. B: Examine a given picture of a spider and describe it.	Text/Tasks: One of A or B. A: Requirement to write both the points and a speech for a debate about zoos. B: Write a response to one of three job advertisements.	Text/Tasks: One of A, B or C. A: Write the text to accompany given photos for a hotel brochure. B: Write a persuasive speech nominating the student of the year. C: Describe the given picture of a house as accurately as possible.
	Section 4: Fiction	
Text: Four short paragraphs from the novel *Robinson Crusoe* (previously unseen text). Tasks: One question requiring retrieval of information, three questions requiring interpretation and inference, and one question requiring students to refer to a short story they studied and to describe aspects of the story's character, location or time period.	Text: Four short paragraphs from the novel *ET: The Extra-Terrestrial* (previously unseen text). Tasks: Two questions requiring inferences about characters, one question requiring students to reflect on human qualities, one question requiring students to comment on the atmosphere of the text, and one question requiring students to refer to a short story they studied and to describe aspects of the story's characters and their relationship with each other, or, aspects of the story which were funny, sad or exciting.	Text: One-and-a-half pages from *Angela's Ashes* by Frank McCourt (previously unseen text). Tasks: One question requiring inference about the character of the teacher in the text, one question requiring students to examine the text for exaggeration as a humorous device and one question requiring students to comment on the suitability of the extract as a basis for a film scene. The second section requires students to refer to a novel or short story they have studied and either comment on the devices used by the author to convey humour or tragedy, or to comment on the author's choice of the title for the novel.

Source: Shiel et al., 2001, Table A6.1.

Further, the emphasis on literary texts and referring to previously studied texts is an aspect of the Junior Certificate English examination that is absent from PISA. (On reviewing more recent Junior Certificate English examination papers,[17] it is apparent that there is little change in the structure of the papers from year to year.)

The higher-level examination consists of two two-and-a-half hour papers, while the ordinary- and foundation-level examinations consist of one two-and-a-half hour paper. Ordinary and foundation level papers are similar in structure, although the ordinary-level paper contains more complex stimulus texts, and a higher proportion of questions which require inference and use of outside knowledge.

Ordinary- and higher-level papers require students to refer to their coursework to a greater degree than the foundation-level paper. Students' responses at higher level are expected to be greater in length and complexity than those at foundation or ordinary levels, and to include aspects of literary criticism and aesthetic appreciation. A broad range of text types and tasks is assessed at all three levels. The balance between course-based texts and unseen texts is achieved by virtue of the fact that students are only required to refer to coursework in half of the sections they attempt.

There are differences between PISA and Junior Certificate English with respect to item format. PISA uses multiple-choice and short open-response formats, whereas the Junior Certificate English Examination requires students to respond to questions with lengthy compositions or commentaries, many of which are literary or expository.

Marking schemes for the Junior Certificate English examination papers[18] are quite descriptive and open to interpretation. Examiners are generally advised to mark by impression. For example, on responses to questions which pertain to the shorter responses for reading comprehension at higher level (as opposed to the marking guides for essay-type responses), examiners are advised to mark on impression and for full marks, "expect candidates to present several points well supported from the text, or fewer points more fully developed" (a guideline which is clearly open to interpretation); however examiners are also supplied with additional broad guidelines, such as that

[17] These are available at http://www.examinations.ie
[18] Again, these are available at http://www.examinations.ie

students may agree or disagree with a particular question, and the location of the relevant pieces of text that they may draw on to respond to the question. They are also provided with exemplar marked responses. Examiners are encouraged to award maximum marks where deserved and to discriminate between those which simply restate points, and those which develop and interpret them. Mechanics of writing contribute little to the overall grade – about 10% regardless of syllabus level. Overall grades according to the marking schemes for higher level are accompanied by the following descriptive standards: 'very good' (high B to A), 'good' (mid C to mid B), 'average' (D to low C), 'poor' (E and lower). The grades are not described in this manner in the guides at ordinary and foundation levels.

The PISA marking guides are more concrete and less impressionistic (PISA Consortium, 2000b; 2003b). This can be related to the item types used and the requirement that student responses be marked in a comparable way in participating countries. In the case of multiple-choice items, the response is by default right or wrong and there is no room for interpretation. In the case of written responses, the PISA marking guides set out clear criteria as to correct and an incorrect responses, giving example correct and incorrect responses for each item. In contrast to the Junior Certificate English examinations, PISA penalises students for reiterating the text rather than addressing the question: a degree of precision in responding is expected in PISA. Further, students are given no credit if their response includes the correct answer but is contradicted by other material they have written. Similar to the Junior Certificate English examination, though, students are not penalised for poor grammar or spelling unless it seriously interferes with the interpretation of what they have written.

Reports from the Chief Examiners of Junior Certificate English for 1994, 2000 and 2003 (available at http://www.examinations.ie) provide additional insights into the strengths and weaknesses of Junior Certificate English candidates. Some common themes are apparent across all syllabus levels. Students commonly misread the question and then provide an inaccurate or incomplete answer. Students are not good at identifying evidence to support conclusions they draw and sometimes fail to elaborate on or justify their choice of response. Students are, by and large, good at creative and narrative writing but overly dependent on these types of writing in answering questions. There is a tendency to summarise when asked to discuss, evaluate, or criticise, with

some evidence of guessing through summarising. There is some evidence of a lack of familiarity or practice with functional texts. Spelling and punctuation is poor. The responses given by some students to some questions on the Junior Certificate English syllabus, therefore, might lack the precision that the open-ended questions in the PISA assessment would require for merit.

2.3.1.2. Junior Certificate Mathematics[19]

The Junior Certificate mathematics syllabus was revised in 2000 and examined for the first time in 2003 (Department of Education and Science/National Council for Curriculum and Assessment, 2000; 2002). Its structure has not changed substantially (Oldham, 2002), although a number of changes have been noted. For examples, there is now no choice on the examination papers, to encourage increased topic coverage; the appropriate use of calculators is recommended, and calculators have been permitted in the examinations since 2003; and geometry has undergone some refinements (see Department of Education and Science/National Council for Curriculum and Assessment, 2002, pp. 3-7).

Concepts are organised into the following topic areas: sets, number systems, applied arithmetic and measure, algebra, statistics, geometry, and functions and graphs. Higher and ordinary level students also study trigonometry. The study of probability (which features in items on the PISA Uncertainty subscale) is reserved for Senior Cycle.

Objectives of the current mathematics syllabus, which apply to all three syllabus levels, may be summarised as follows:

A. Recall of mathematical facts
B. Instrumental understanding
C. Relational understanding
D. Application of mathematical knowledge
E. Analysis of information, including that presented in unfamiliar contexts
F. Ability to create mathematics for oneself (e.g., make informed guesses)
G. Development of psychomotor skills to attain objectives
H. Ability to communicate mathematics

[19] This review of the Junior Certificate mathematics syllabus and examinations is largely based on the review of Cosgrove et al. (2001), supplemented where appropriate with more recent commentaries and analyses.

I. Appreciation of mathematics

J. Awareness of the history of mathematics.

Of the ten objectives, six (A, B, C, D, G and H) are assessment objectives, examined through the Junior Certificate mathematics examination, while the remaining four are not.

The rationale provided for each syllabus level indicates there are differences in the extent to which students are expected to apply mathematical concepts and demonstrate understanding in a variety of contexts (Department of Education and Science/National Council for Curriculum and Assessment, 2000). At higher level, the syllabus is geared towards students who are of above average mathematical ability, some of whom will use academic mathematics in the future; therefore a balance must be struck between challenging the most able students and encouraging those who are developing at a slightly slower pace and the development of abstraction and generalisation skills is emphasised alongside the introduction of proofs. Ordinary level is geared towards average ability students and offers mathematics that is both meaningful and accessible, providing for the gradual introduction of more abstract ideas. The emphasis is on the development of mathematics as a body of knowledge and skills that make sense and that can be used in many different ways. The foundation-level course objectives involve developing knowledge and skills in basic mathematics and awareness of the usefulness of mathematics. The emphasis is on building confidence, both in the students themselves, and in their involvement with mathematics as a discipline.

Cosgrove et al. (2005, pp. 166-167) have compared the content of the mathematics course at the three syllabus levels and comment that while there is not a substantial difference in topics covered on the higher- and ordinary-level courses, the foundation-level course focuses more on the types of mathematical concepts and operations that one is likely to encounter in everyday life (such as those involving money, percentages, area). The main difference between higher and ordinary level courses is in terms of the depth of topic coverage and the extent to which students are required to be familiar with theorems and proofs.

A comparison of the aims and objectives of the Junior Certificate mathematics curriculum and the PISA mathematics assessment, and of the PISA test items and Junior

Certificate examination papers, indicates a substantial divergence in what is learned and assessed. This can be traced first to the fact that the Junior Certificate mathematics assessment objectives are likely to take higher priority than objectives which are not assessed (Cosgrove et al., 2005). For example, the real-life approach to mathematical problem-solving in PISA implies that the ability to solve problems in novel, authentic contexts is an important prerequisite for many of the items (97% of items were rated as having a real-life context in the case of the PISA 2000 mathematics item set) (see Chapter 1). This skill is not apparent in any of the assessment objectives, although it is mentioned in the Objective E (which is not assessed). In the Junior Certificate, questions are usually presented in a purely mathematical and abstract context, almost always without redundant information. In the PISA assessment, questions are usually embedded in rich real-life contexts (Cosgrove et al., 2005). Second, Junior Certificate mathematics emphasises vertical mathematisation (developing increasingly complex mathematics concepts and skills in abstract contexts) (Oldham, 2002). PISA, in contrast, emphasises horizontal mathematisation (the application of mathematical concepts and skills to organise and solve a problem located in a real-life situation, and the abstraction of concepts and skills from these contexts) (OECD, 2003b; Treffers, 1987). Third, it has been suggested that Junior Certificate mathematics tends not to tap processes associated with items in the PISA Reflection cluster (Cosgrove et al., 2005). The structure of the Junior Certificate mathematics examination further demonstrates the differences in the relative emphasis on the application of memorised procedures and problem-solving in novel contexts. Questions on the Junior Certificate mathematics examination are typically divided into three parts (a, b, c). Questions in part a test recall and/or simple instrumental understanding; part b generally tests procedures involving instrumental understanding with which students should be familiar; it can also assess relational understanding; and part c is intended to address somewhat higher-order objectives, but still in fairly familiar contexts. Credit is given to parts a, b and c at a ratio of 1:2:1. Close and Oldham (2005) comment that part a is intended to ease students into the question, and part b to reward diligent learning. Only an almost perfect performance on all parts a and b would result in a mark of 55% (necessary for a grade C). For B and especially A grades, though, students *should* have to display the higher-order skills associated with application and problem-solving in comparatively familiar contexts: questions in part c were intended to be to a certain extent unpredictable, but in practice, they have tended not to be. Close and Oldham comment that, with regard to the role of

non-mathematical (realistic) contexts, there were concerns that these would not provide a "level playing field" for candidates. The Junior Certificate mathematics teacher guidelines also emphasise the difficulty in assessing higher-order skills in realistic contexts in the time and other constraints imposed by the conditions under which the Junior Certificate mathematics examination is attempted. However, this runs the risk that part c problems become amenable to reduction from application or problem-solving status to the status of rehearsed procedures. This may reflect inherently procedural views of mathematics (Lyons et al., 2003; Oldham 2001, 2003; see also the observations documented in Chief Examiners' reports described in the following paragraph). Fourth, it has also been noted (Cosgrove et al., 2005), from a comparison of the PISA mathematics marking schemes (PISA consortium, 2000a; 2003a) that the approach to marking mathematics in the Junior Certificate Examination offers greater scope than the PISA assessment for recognising merit in students' work. The marking schemes for PISA treat most questions, even those that require extensive working out and justification/explanation of the solution, as right or wrong (although partial credit is applied to some items to distinguish between complete and incomplete working out). The marking schemes for Junior Certificate mathematics are more detailed, and a zero mark for a question is much less common. This is because questions are presented in units, each of which is allocated a maximum mark, typically 5 or 10 marks. Each line of the student's work is scrutinised and subjected to penalties (e.g., one mark may be deducted for an arithmetical slip and three for a more serious error such as a misapplication of an algebraic rule). The same error is penalised once only in any one part of a question. In the application of these penalties, a student's mark is not allowed to drop below the 'attempt mark' for that part which is usually one third of the maximum mark for the section. Therefore, a student who makes a worthwhile attempt at a question with a maximum mark of 10 will receive at least 3 marks. Close and Oldham (2005) have commented that the placement of the Junior Certificate marking schemes in the public domain serves a valuable transparency function, but also runs the risk of a mark-focused approach to learning and instruction.

The 1999 and 2003 Chief Examiners' Reports on Junior Certificate Mathematics (available at http://www.examinations.ie) indicate that many students appear to approach Junior Certificate mathematics in a mechanical manner, are not using higher-order reasoning in working out/checking their answers, and that some fundamental

conceptual understanding is lacking. Aspects of geometry, algebra and trigonometry were identified as general areas of weakness. In contrast, students typically performed well on questions that called for the application of basic concepts involving number, applied arithmetic, statistics and functions. The ability of students to lay out their responses in a neat and methodical manner was noted.

Overall, these observations suggest that students who approach mathematics at a mechanical level will find PISA mathematics extremely challenging; that students are ill-prepared for the manner in which PISA mathematics problems are contextualised; and that students will also be relatively unfamiliar with the task demands of PISA test items which are in the Reflect competency cluster.

2.3.2. Quantitative Comparisons: the Test-Curriculum Rating Project

This section outlines the framework of the test-curriculum rating project implemented in Ireland, describes the rating scales used, how the method represents a development from the TIMSS test-curriculum matching analysis, and considers the methodological and conceptual limitations associated with the test-curriculum rating project.

2.3.2.1. Analysis of Curriculum Coverage: An Example from TIMSS 1995

The approach used in TIMSS 1995 to measure OTL is similar in some respects to that adopted for analyses of the curricula in Ireland (Shiel et al., 2001; Cosgrove et al., 2005) and is described here in order to demonstrate how the analyses of Irish curricula build on the TIMSS approach. In TIMSS 1995, the distinction was made between intended, implemented and attained curriculum (as in the Second International Mathematics Study (SIMS; Robitaille & Garden, 1989)). Beaton et al. (1996a) distinguished between these three components as follows:

> The **intended curriculum** is composed of the mathematics and science instruction and learning goals defined at the system level. The **implemented curriculum** is the mathematics and science curriculum as interpreted by teachers and made available to students. The **attained curriculum** is the mathematics and science content that students have learned and their attitudes towards these subjects. (pp. A1-A2, bold type in original)

The aim of the TIMSS 1995 TCMA (e.g., Beaton et al., 1996a) was to examine the effect of topic inclusion/exclusion on national *intended* curricula on student

achievements. TIMSS obtained item-level data by asking mathematics and science curriculum experts in each country to rate each item as to whether or not the topic covered by the item was included in their intended curriculum or not. Thus a dichotomous variable was associated with each item in each country and it was possible to calculate and compare percent correct of all TIMSS items with percent correct of only those TIMSS items which, according to national curriculum experts, were covered in the intended national curricula. Separate analyses were carried out for each of the four TIMSS grade levels. Broadly speaking, results indicated that the TIMSS assessments were seen to be a fair test and largely appropriate to national curricular aims in the majority of countries. Beaton et al. (1996a, p. B5) concluded: "It is clear that the selection of items does not have a major effect on the general relationship [of achievement] among countries". For example, regarding the mathematics items, at second year level, curriculum experts in Ireland deemed 89% of items appropriate, and there was virtually no difference between the percent correct on these items only compared with the percent correct on all mathematics items (58% and 59%, respectively). At first year level, 70% of items were seen to be appropriate; again; there was little difference between percent correct for these items only (55%) compared with all mathematics items (53%).

2.3.2.2. Limitations of the TIMSS Approach to Assessing Curriculum Coverage

There are some limitations to the methodology used in TIMSS 1995. First, using a dichotomous variable to indicate topic coverage/no topic coverage is overly simplistic. It does not account for the possibility that students might be familiar with some characteristics of an item (for example, the underlying concept) but not familiar with other characteristics of an item (for example, the context in which the concept is applied) (see Floden, 2002, p. 241). Nor does it allow for the fact that students may have had the opportunity to learn a topic in a broad, but not a detailed, manner, i.e., that there might be gradations of familiarity. A polytomous and multidimensional rather than a dichotomous unidimensional indication of curricular coverage might better capture complex differences between items and between and within countries.

Second, using the same rating for all students at each grade level does not account for differences in curricular coverage which are dependent on academic track. Differences by academic track were not examined in TIMSS until TIMSS 1999, and also in TIMSS

2003 (Martin et al., 2000a; Mullis et al., 2000; Martin et al., 2004; Mullis et al., 2004). The Irish education system is notable in that while the vast majority of students study the same programme (the Junior Certificate), which is classified as an academically-oriented programme on the International Standard Classification of Education (ISCED; OECD, 1999a), there is, as we have seen from the descriptions of Junior Certificate English and mathematics, marked differences between syllabus levels in the level of complexity subjects are taught for both English and mathematics. Thus, in Ireland at least, a measure of curricular coverage should be not only multidimensional with respect to the properties of items, but should also be multidimensional with respect to syllabus level or academic track.

2.3.2.3. Framework for the Irish Test-Curriculum Rating Project

The framework for the test-curriculum rating project, conducted in both 2000 and 2003 to supplement analyses of the PISA data in Ireland, comprises a 3 x 3 matrix whereby the three aspects of the items which are of interest are cross-classified with the three syllabus levels. In English/reading, process, context/application, and format were examined, while in mathematics, concept, context/application, and format were examined.[20] Ratings range from 1 ('not familiar') to 3 ('very familiar') (Table 2.3). The scales and the manner in which ratings were applied are described in more detail in Shiel et al. (2001, pp. 224-232) and Cosgrove et al. (2005, pp. 269-270).

Six individuals with extensive knowledge of the curriculum area in question (mathematics or English) and/or teaching experience at post-primary level assigned ratings to the items (three raters for mathematics and three for English/reading). Each individual was briefed as to the nature of the task and provided with a copy of the materials needed to rate the items (the PISA assessment framework for the domain in question; a sample test item with ratings and a rationale and explanation for the ratings assigned; a detailed description of the rating scales and how to apply them; syllabus documents and teacher guidelines for the subject in question; and copies of recent Junior Certificate examination papers). Initially, items were rated independently, and items on which there was a lack of consensus were flagged. 'Consensus' was defined as the modal rating assigned to a particular scale at a particular syllabus level where there

[20] These particular aspects were selected in the course of a series of planning and developmental meetings with curriculum experts in each subject area.

was either perfect agreement across the three raters *or* where there was disagreement, the difference did not exceed one scale point. Consensus was reached on the flagged items during a meeting with the raters.

Table 2.3. Framework for the Test-Curriculum Rating Project

Subject/Aspect	*Junior Certificate Syllabus Level*		
English/Reading	*Higher*	*Ordinary*	*Foundation*
Process: How familiar would you expect the typical third year student to be with the specific reading process(es) underlying this item?	Not/Somewhat/ Very Familiar	Not/Somewhat/ Very Familiar	Not/Somewhat/ Very Familiar
Context/Application: How familiar would you expect the typical third year student to be with the application of the specific reading process(es) underlying this item in the type of context (genre, text length, density, complexity) suggested by the item and stimulus text?	Not/Somewhat/ Very Familiar	Not/Somewhat/ Very Familiar	Not/Somewhat/ Very Familiar
Format: How familiar would you expect the typical third year student to be with the application of the specific reading process(es) underlying this item in the type of format suggested by the item and stimulus text?	Not/Somewhat/ Very Familiar	Not/Somewhat/ Very Familiar	Not/Somewhat/ Very Familiar
Mathematics			
Concept: How familiar would you expect the typical third year student to be with the specific mathematical concept(s) underlying this item?	Not/Somewhat/ Very Familiar	Not/Somewhat/ Very Familiar	Not/Somewhat/ Very Familiar
Context/Application: How familiar would you expect the typical third year student to be with the application of the specific mathematical concept(s) underlying this item in the type of context suggested by the item and stimulus text?	Not/Somewhat/ Very Familiar	Not/Somewhat/ Very Familiar	Not/Somewhat/ Very Familiar
Format: How familiar would you expect the typical third year student to be with the application of the specific mathematical concept(s) underlying this item in the type of format suggested by the item and stimulus text?	Not/Somewhat/ Very Familiar	Not/Somewhat/ Very Familiar	Not/Somewhat/ Very Familiar

Source. Shiel et al., 2001, Tables 6.15 and 6.16.

In PISA 2003, in addition to supplying the familiarity ratings, the mathematics panel examined the concept underlying each PISA mathematics item and identified the mathematical topic area on the Junior Certificate in which the concept was most likely to be.

2.3.2.4. Results of the Irish Test-Curriculum Rating Project

Results from this section are summaries of analyses reported in Chapter 6 of the PISA 2000 national report (Shiel et al., 2001) and from Chapter 6 of the PISA 2003 national report (Cosgrove et al., 2005).

2.3.2.4.1. PISA 2000 reading

At higher and ordinary levels, the process underlying the PISA reading items was rated as 'somewhat familiar' or 'very familiar' in over 90% of cases. At foundation level, a somewhat lower percentage of items (75%) was rated as 'somewhat familiar' or 'very familiar' (Table 2.4). Differences between syllabus levels are more marked for the context of application ratings, where abut half the items were judged to be not familiar at foundation level (compared with 18% at ordinary level and 13% at higher). Familiarity with item format was considerably lower than for process; however this arises largely due to the fact that the Junior Certificate English examination papers do not include multiple-choice or short response item format. This is not to say that students would not have encountered these item formats in other contexts.

Table 2.4. PISA 2000 Reading Curriculum Familiarity Ratings, by Junior Certificate Syllabus Level

	Not familiar		*Somewhat familiar*		*Very familiar*	
	N	*%*	*N*	*%*	*N*	*%*
Process						
Higher	5	3.7	21	14.7	115	81.6
Ordinary	14	9.6	52	36.8	76	53.7
Foundation	35	25	66	47.1	39	27.9
Context						
Higher	19	13.2	36	25.7	86	61
Ordinary	26	18.4	77	54.4	38	27.2
Foundation	71	50.7	66	47.1	3	2.2
Format						
Higher	71	50	22	15.4	49	34.6
Ordinary	74	52.2	33	23.5	34	24.3
Foundation	102	72.1	32	22.8	7	5.1

Source: Shiel et al., 2001, Table 6.18.

In sum, it would appear that there is considerable overlap between PISA reading and Junior Certificate English in terms of the reading processes assessed. The context in which these processes are applied is expected to be unfamiliar to foundation-level students in the case of about half of the items (where the text genre and density of the text were expected to be outside the scope of study of these students). The relatively low familiarity associated with item format arises due to the preponderance of longer, essay-type responses required in the Junior Certificate English examinations.

Table 2.5 shows the percentage of items rated 'not familiar' on the process scale for each syllabus level, by the three PISA processes and two PISA text types.[21] At higher level, processes underlying around 90% or more of items are expected to be at least somewhat familiar to students, for ordinary level, this applies to about 88% or more of items. At foundation level, students were expected to be familiar with between three-fifths and five-sixths of items depending on the reading process assessed. Across all syllabus levels, students were expected to be less familiar with items associated with the reflect subscale. Comparing the process ratings by text type, expected familiarity is greater for continuous texts compared with non-continuous ones for all three syllabus levels.

Table 2.5. *Percent of PISA 2000 Reading Items Rated 'Not Familiar' on Process, by Process Area and Text Type, for Higher, Ordinary and Foundation Level*

	Higher	*Ordinary*	*Foundation*
Process Area			
Retrieve	1.7	5.0	20.0
Interpret	0.0	5.0	17.5
Reflect	11.1	22.2	41.7
Text Type			
Continuous	1.1	2.2	13.3
Non-Continuous	8.7	23.9	47.8

Taking into account both the English syllabus level studied by each student for the Junior Certificate and the particular set of items the student attempted during the PISA 2000 assessment, the average expected familiarity of each student with the set of items attempted was computed for each of the three aspects. Pearson correlations were then computed to assess the degree of association between student familiarity and achievement on PISA 2000 reading. Correlations were all significant and strongest for process at .55; the correlations for context and format were .54 and .46, respectively. Given that the scales are logically interdependent (whereby if process was rated unfamiliar, then so was context of application and concept), a composite scale was constructed. The correlation between the composite scale and reading achievement was .56. Analyses of curriculum familiarity were not carried out as part of PISA 2003 since no new reading items appeared in PISA 2003.

[21] These analyses were not published previously but are included here for the sake of consistency with the PISA 2003 mathematics analyses.

2.3.2.4.2. PISA 2003 mathematics

Table 2.6 shows the curriculum familiarity ratings for PISA 2003 mathematics for each of the three item aspects. Concepts underlying the majority of items at higher (69%) and ordinary (65%) levels were somewhat or very familiar, while just under half of the items at foundation level (48%) were rated as somewhat or very familiar. In contrast, the contexts in which the mathematics problems were presented were rated as unfamiliar in the majority of items (66% at higher level, 71% at ordinary level, and 80% at foundation level). Item formats were also largely unfamiliar to Irish students, regardless of syllabus level (at least in the context of Junior Certificate mathematics, although as with English/reading, students may encounter multiple-choice item formats in other contexts). Despite the addition of two new mathematical areas (Quantity, Uncertainty) to the PISA 2003 assessment (PISA 2000 assessed Space & Shape and Change & Relationships only), the ratings for PISA 2003 are very similar to those given in PISA 2000.

Table 2.6. PISA 2003 Mathematics Curriculum Familiarity Ratings, by Junior Certificate Syllabus Level

	Not familiar		*Somewhat familiar*		*Very familiar*	
	N	*%*	*N*	*%*	*N*	*%*
Concept						
Higher	26	30.6	21	24.7	38	44.7
Ordinary	30	35.3	25	29.4	30	35.3
Foundation	44	51.8	22	25.9	19	22.4
Context						
Higher	56	65.9	19	22.4	10	11.8
Ordinary	60	70.6	17	20.0	8	9.4
Foundation	68	80.0	14	16.5	3	3.5
Format						
Higher	53	62.4	21	24.7	11	12.9
Ordinary	62	72.9	17	20.0	6	7.1
Foundation	71	83.5	12	14.1	2	2.4

Source: Cosgrove et al., 2005, Table 6.13.

Concept familiarity ratings were also compared both for the PISA subscale areas and for the three PISA competency clusters (Table 2.7).

Table 2.7. Percent of PISA 2003 Mathematics Items Rated 'Not Familiar' on Concept, by Content Area and Competency Cluster, for Higher, Ordinary and Foundation Level

	Higher	*Ordinary*	*Foundation*
Content Area			
Space and Shape	30.0	35.0	50.0
Change and Relationships	22.7	27.3	50.0
Quantity	26.1	26.1	39.1
Uncertainty	45.0	55.0	70.0
Competency			
Reproduction	19.2	23.1	38.5
Connections	35.0	37.5	50.0
Reflection	36.8	47.4	73.7

Source: Cosgrove et al., 2005, Tables 6.14 and 6.15

Irish students were expected to be familiar with the concepts underlying the majority of the items on the Quantity subscale (74% at higher and ordinary levels; 61% at foundation level). Familiarity with concepts underlying the Change & Relationships subscale was higher for higher and ordinary levels compared with foundation levels (77% at higher, 73% at ordinary, 50% at foundation). Ratings on the Space & Shape items suggest moderate familiarity, while students were expected to be least familiar with items on the Uncertainty subscale. The majority of Reproduction items (62% to 80% depending on syllabus level) are expected to be somewhat or very familiar to students at all syllabus levels, while ratings on the Connections items suggest moderate familiarity (50% to 65%). Reflection items are somewhat less familiar to students, particularly at foundation level, where 74% of such items were rated as being unfamiliar.

Student-level familiarity ratings were computed in the same manner as for English/reading in PISA 2000. A single composite curriculum familiarity scale, incorporating all three aspects, was also created. The correlation between concept familiarity and mathematics achievement, at .37, is the highest; item format correlates .28 with achievement, and context correlates .21. (The respective correlations for PISA 2000 mathematics are .48, .23 and .20.) The composite curriculum familiarity scale has a correlation of .32 with combined mathematics achievement. These findings suggest that concept familiarity is most strongly predictive of success on an item.

Concepts underlying mathematics items were also classified according to which Junior Certificate mathematics topic area they best fit. This classification indicated that the Junior Certificate mathematics topic areas of sets, geometry, and trigonometry are not assessed at all by the PISA mathematics items. There is also little coverage in PISA of algebra, and functions and graphs. The majority of PISA mathematics items whose concepts *are* somewhat familiar to Irish students are located in the Junior Certificate mathematics topic areas of applied arithmetic and measure, and statistics (Table 2.8).

Table 2.8. Curriculum Area Ratings for PISA 2003 Mathematics Items Cross-tabulated with Junior Certificate Mathematics Syllabus Level

	Junior Certificate mathematics strand area									
	Not in Junior Cert.		*Number systems*		*Applied arith. & measure*		*Algebra*		*Statistics*	
Syllabus Level	N	%	N	%	N	%	N	%	N	%
Higher	26	28.6	8	8.8	30	33.0	5	5.5	18	19.8
Ordinary	30	33.0	9	9.9	29	31.9	4	4.4	16	17.6
Foundation	44	49.4	8	9.0	23	25.8	1	1.1	13	14.6
	Functions and graphs		*Sets*		*Geometry*		*Trigonometry*		*Total*	
Syllabus Level	N	%	N	%	N	%	N	%	N	%
Higher	4	4.4	0	0.0	0	0.0	0	0.0	91	100.0
Ordinary	3	3.3	0	0.0	0	0.0	0	0.0	91	100.0
Foundation	0	0.0	0	0.0	0	0.0	n/a	n/a	89	100.0

Note. Total number of PISA 2003 mathematics items = 85. As evidenced in the totals, 6 items were identified as being located in two Junior Certificate strand areas in the case of higher and ordinary levels, and 4 items in the case of foundation level.

Source: Cosgrove et al., 2005, Table 6.10.

Cosgrove et al. (2005) also compared the ratings in Table 2.8 for each PISA mathematics content area. Results indicated that concepts underpinning the PISA items were, at times, distributed across several Junior Certificate mathematics topic areas. At higher level for example, concepts underlying the 18 items associated with PISA Change & Relationships which are on the Junior Certificate syllabus are spread across five Junior Certificate topic areas (number systems, applied arithmetic and measure, algebra, statistics, and functions and graphs). Items associated with PISA Quantity are spread across three Junior Certificate areas (number systems, applied arithmetic and measure, and functions and graphs). The PISA Space & Shape items rated as somewhat or very familiar in terms of their underlying concept, which one might expect to be associated with the topic area of geometry, were almost all located in the Junior Certificate topic area of applied arithmetic and measure. Almost all PISA Uncertainty items were located in the Junior Certificate topic area of statistics.

2.3.2.5. Corroborating Evidence for the Results of the PISA 2003 Test-Curriculum Project for Mathematics

Close and Oldham (2005), as a complement to the analyses of PISA mathematics and Junior Certificate mathematics (Cosgrove et al., 2005), analysed the questions on the 2003 Junior Certificate mathematics examination papers with respect to the PISA mathematics assessment framework. That is, while Cosgrove et al. mapped PISA onto the Junior Certificate, Close and Oldham mapped the Junior Certificate onto PISA. The PISA mathematics framework was used as a guide in this mapping process and all items were 'forced' into the three aspects of the PISA framework (overarching idea, competency cluster, and context). Results provide corroborative evidence for the findings of Cosgrove et al., and provide further insights into how the two assessments differ.

Close and Oldham (2005) found that, while, in PISA, there are approximately equal percentages of items assessing the four overarching ideas, in the 2003 Junior Certificate papers, the percentages of items assessing Quantity range from 17% at higher level, 32% at ordinary level, and 53% at foundation level. While 34% of higher-level questions were classified as Space & Shape, this figure is 27% at ordinary level and 13% at foundation level. Change & Relationships also shows a pattern of reduced emphasis across the syllabus levels (38%, 32%, and 22%). In the case of Uncertainty there are considerably more items in the PISA tests than in the Junior Certificate papers; the figures range from 10-13% for the Junior Certificate. Regarding the competency clusters, in PISA, 31% of items assess Reproduction, 47% Connections, and 22% Reflection. Close and Oldham reported that there are no items at all in the Reflection cluster on any of the 2003 Junior Certificate mathematics examination papers. All foundation-level items were classified as Reproduction questions, while 93% and 83% of ordinary- and higher-level questions, respectively, were classed as Reproduction questions. Close and Oldham also found that, across the three Junior Certificate papers the mean percentage of mathematics items in realistic situations is 33% for the Junior Certificate compared with 80% in PISA.

2.3.2.6. Limitations of the Irish Test-Curriculum Rating Project

The test-curriculum rating project provides a framework for discussing similarities and differences between the PISA approach to assessing achievement and what the State examinations assess at the end of Junior Cycle. However, the analyses suffer from several limitations. First, the test-curriculum rating project does not take into account the likelihood that numerous factors, other than curriculum intent and the manner in which it is implemented, affect student achievements (the attained curriculum). Indeed, there are additional characteristics of the PISA assessment which are relevant to this analysis, such as the manner in which students' responses are marked or graded, which have not been included in the analysis.

Second, the analysis does not give detailed information on which elements of the Junior Certificate Examination are not assessed by PISA in the case of reading. It was noted, however, from qualitative comparisons of the two assessments, that PISA reading assesses little if any of the oral/aural and writing strands. Further, the bulk of texts studied for Junior Certificate English comprise literary pieces (including novels, short stories, plays, and poetry). The PISA reading test included only two texts which might be described as literary: one short story (*The Gift*) and one poem (*If*). It was also noted that the relative emphasis in PISA on non-continuous functional texts is higher than in Junior Certificate English (if the content of Junior Certificate English examination papers are representative of the types of materials students study in preparation for the examination). In PISA 2003, an attempt to quantify this issue was made with respect to mathematics. Each mathematics item was mapped onto the Junior Certificate mathematics syllabus and this revealed considerable disparities between the two assessments, whereby up to half of the Junior Certificate mathematics topics are not assessed in PISA.

Third, the results are only interpretable in a national context so no conclusions may be drawn about the relative level of expected familiarity with the assessments. For example, the low familiarity of Irish students with PISA 2003 Uncertainty items may not be so low, relatively speaking, in an international context.

Fourth, student ability and curriculum familiarity may be confounded. If one considers the results of two ordinary-least-squares regressions (one for PISA 2000 reading, one

for PISA 2003 mathematics) which examine the association of familiarity with process (in the case of reading) or concept (in the case of mathematics) after adjusting for performance on the relevant Junior Certificate subject (not the most appropriate adjustment for ability, but the only available one), the association between curriculum familiarity and performance on PISA is much weaker (borderline significant in the case of reading, and not significant in the case of mathematics) (Table 2.9).[22] While a more generic ability measure on intake would have been preferable in these analyses, results are nonetheless suggestive of the limited explanatory power of these data. This may be attributable to the broad nature of the ratings and the fact that they are related to student achievement in a static manner, rather within a longitudinal framework, which could measure relative gains over time. Floden (2002) has suggested that OTL measures may in essence be confounded with the types (and levels of complexity) of skills and concepts to which students are exposed; that there is an endogenous quality to these measures. Other researchers have found weak or inconsistent associations between OTL measures and achievement (e.g., Floden, 2002; Lapointe et al., 1989, p. 33; Lapointe et al., 1992, pp. 31-39).

Table 2.9. *Ordinary-Least-Squares Regression with Achievement on PISA Reading (2000) and PISA Mathematics (2003) as the Outcome Variables, with Familiarity with Process and Performance on the Junior Certificate (2000)/Familiarity with Concept and Performance on the Junior Certificate (2003) as Explanatory Variables*

	PISA 2000 Reading				*PISA 2003 Mathematics*		
	r	*t*	*p*		*r*	*t*	*p*
Process	.033	1.980	.051	Concept	.030	1.387	.169
JCE English	.739	59.943	<.001	JCE Mathematics	.698	37.297	<.001

2.3.3. Comparisons of Performance on PISA and the Junior Certificate

2.3.3.1. Comparison of Achievements on the Two Assessments

The achievement results for Ireland for PISA 2000 and 2003 have been published in several reports (e.g., Cosgrove et al., 2005; OECD, 2001b; 2004c; Kirsch et al., 202; Shiel et al., 2001) and are considered here only to provide a broad description of Irish performance in international terms, and a context in which to interpret performance on the Junior Certificate.

[22] These data were previously unpublished but are shown here to provide support for the argument that the ratings confound ability with familiarity.

2.3.3.1.1. Review of results for Ireland – PISA 2000 reading

By international standards, Irish achievement on the PISA 2000 test of reading is high overall, with comparatively fewer weak readers, and comparatively more students with proficient or advanced levels of reading.

Ireland's mean scores on the combined reading scale, on the three process subscales, and on the two text format subscales are all significantly higher than the corresponding OECD country average scores (which were around 500). Ireland achieved the fifth highest mean score (526.7) among the 27 OECD countries that met agreed criteria on school and student participation levels.[23] The performance of Irish students on the Retrieve (524.3) and Interpret (526.5) subscales is about the same as on the test as a whole. Ireland ranked third on the Reflect subscale, with a mean score (533.2) that does not differ significantly from that of Canada, the highest scoring country on the subscale. Ireland ranked fourth on the continuous text subscale, with a mean score of 528, and sixth on the non-continuous text subscale, with a mean score of 530.

In Ireland, 3.1% of all students are below proficiency Level 1 (compared to an OECD average of 6.0%); 7.9% are at Level 1 (compared to an OECD average of 11.9%). At the upper end of the scale, 41.3% of students were at Levels 4 (27.1%) and 5 (14.2%) (the corresponding OECD averages are 22.3% and 9.5%, respectively). The distribution of achievement across proficiency levels was similar for the three subscales. A comparison of the scores of students in Ireland at the 10th and 90th percentile points indicates a 35-point difference at the 10th percentile (401.3 compared to 365.9 across the OECD) and an 18-point difference at the 90th percentile (641.1 compared to 622.7).

The standard deviation associated with the combined reading scale (93.6) is somewhat smaller than the OECD average, indicating comparatively narrow dispersion in achievement scores. The comparatively low between-school variance (discussed in more detail in Section 2.4) suggests that schools are comparatively homogenous with respect to achievement and that the majority of achievement differences lie within schools (students and classes), rather than between them.

[23] One country, the Netherlands, was not included in reports of achievement since its response rates were too low to ensure the reliability of the sample.

2.3.3.1.2. Review of results for Ireland – PISA 2003 mathematics

By international standards, Irish achievement on the PISA 2003 mathematics assessment is around the OECD average. The distribution of performance is relatively homogenous, characterised by fewer high achievers, as well as fewer low achievers.

The mean score for Ireland on the combined mathematics scale (502.8) does not differ from the OECD average (500.0). In contrast to reading, where performance was strong on all subscales, there is some variability in the mean performance of students. Irish performance was weakest on the Shape & Space subscale, where it is significantly below the average (476.2 compared with 496.3). The mean score for Ireland on the Change & Relationships scale is significantly above the OECD average, albeit by just 7 points (506.0 compared with 498.8). Performance on the Quantity subscale does not differ from the OECD average (501.7 compared with 500.7). Performance on the Uncertainty subscale was highest, and significantly above the OECD average (517.2 compared to 502.0).

In Ireland, 4.7% of all students are below proficiency Level 1 (compared to an OECD average of 8.2%); 12.1% are at Level 1 (compared to an OECD average of 13.2%). At the upper end of the scale, 31.5% of students were at Levels 4 to 6 – 20.2% at level 4, 9.1% at Level 5, and just 2.2% at Level 6. The OECD average at Levels 4 to 6 is 33.7% (19.1% at Level 4, 10.6% at Level 5, and 4.0% at Level 6). The distribution of achievement across proficiency levels is broadly similar for the four subscales. A comparison of the scores of students in Ireland at the 10th and 90th percentile points of the combined mathematics scale indicates a 41-point difference at the 10th percentile (393.1 compared to 351.9 across the OECD) and an 14-point difference at the 90th percentile (613.9 compared to 628.3 across the OECD).

Similar to the results for PISA 2000 reading, the standard deviation associated with the combined mathematics scale (85.3) is smaller than the OECD average (98.6), indicating comparatively narrow dispersion in achievement scores. Comparatively low between-school variance (discussed in Section 2.4) is also evident in the achievement variance associated with the Irish results.

2.3.3.1.3. Performance on PISA at higher, ordinary and foundation levels

The national reports for Ireland for both PISA 2000 and PISA 2003 recorded large differences in performance between students taking Junior Certificate English and mathematics at higher, ordinary and foundation levels. The mean scores of students on PISA 2000 reading taking Junior Certificate English at higher, ordinary and foundation, respectively, are 562 (SE = 2.1), 451 (SE = 3.9) and 336 (SE = 9.8). These correspond, respectively, to the 62nd, 20th, and 3rd percentiles for Ireland. The respective scores of students on PISA 2003 mathematics at the three levels of the Junior Certificate mathematics examination are 563 (SE = 2.1), 469 (SE = 2.0) and 385 (SE = 5.2). Respectively, these correspond to the 76th, 34th, and 8th national percentiles.

Performance differences are also evident when one considers the percentages of students taking each syllabus level at each PISA proficiency level. Figure 2.1 and Table 2.10 show the distribution of students across each PISA combined reading proficiency level for the 2000 survey, by Junior Certificate English syllabus level. There is considerable disparity in the distribution of achievement across syllabus levels. At foundation level, around 90% of students are at or below Level 1; about 51% are below Level 1, indicating that the reading literacy skills of half of these students are not reliably assessed by PISA. Only around 10% of foundation-level students score at Levels 2 or 3, and none at Levels 4 or 5. At ordinary level, a sizeable minority of students score at (21.3%) or below (7.4%) Level 1. The modal proficiency level at ordinary level is Level 2 (achieved by around 35% of students), while about 27% achieve at Level 3. Just under 9% score at Level 4, and 1% at Level 5. The modal proficiency level for higher-level students is Level 4 (achieved by 35%); a further 20% achieve Level 5. Just over 2% achieve at or below Level 1.

Figure 2.1. Distribution of Students Taking English at Higher, Ordinary and Foundation Levels Across PISA Reading Proficiency Levels: PISA 2000

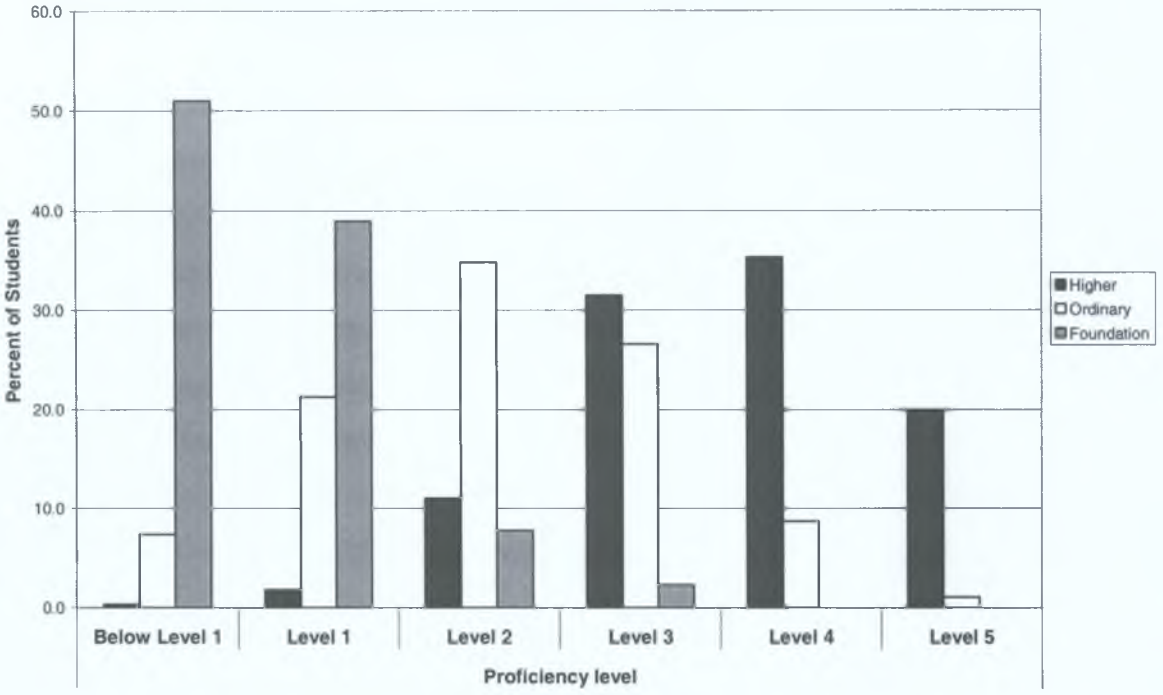


Table 2.10. Distribution of Students Taking English at Higher, Ordinary and Foundation Levels Across PISA Reading Proficiency Levels: PISA 2000

	Below Level 1		Level 1		Level 2		Level 3		Level 4		Level 5	
Syllabus	*%*	*SE*	*%*	*SE*	*%*	*SE*	*%*	*SE*	*%*	*SE*	*%*	*SE*
Higher	0.4	0.15	1.8	0.36	11.0	0.91	31.5	1.15	35.3	1.27	20.0	1.00
Ordinary	7.4	1.24	21.3	2.10	34.8	2.03	26.6	2.42	8.7	1.18	1.1	1.43
Foundation	51.0	9.54	39.0	8.26	7.8	6.22	2.3	2.85	0.0	0.00	0.0	0.00

Source: Cosgrove et al., 2005, Table 6.20

Figure 2.2 and Table 2.11 show the distribution of students across each PISA combined mathematics proficiency level for 2003, by Junior Certificate mathematics syllabus level. Again, there is substantial disparity in the distribution of achievement across syllabus levels. At foundation level, around 72% of students are at or below Level 1; and about one in three is below Level 1, indicating that the mathematics skills of these students are not reliably assessed by PISA. Just under 23% of foundation-level students score at Level 2 and just 5.5% at Level 3; no foundation-level students achieved above Level 3. At ordinary level, a sizeable minority of students score at (17.8%) or below (4.1%) Level 1. The majority of ordinary-level students score at Levels 2 (36.2%) or 3 (30.4%). One in ten ordinary-level students scored at Level 4, and a minority – about 1.6% – demonstrated the more advanced mathematics skills associated with Levels 5 and 6. At higher level, just 1.5% of students scored at or below Level 1, 9% at Level 2, and about 29% at Level 3. The modal proficiency level for higher-level students is Level 4 (achieved by 36%); a further 25% achieve Levels 5 and 6.

Figure 2.2. Distribution of Students Taking Mathematics at Higher, Ordinary and Foundation Levels Across PISA Mathematics Proficiency Levels: PISA 2003

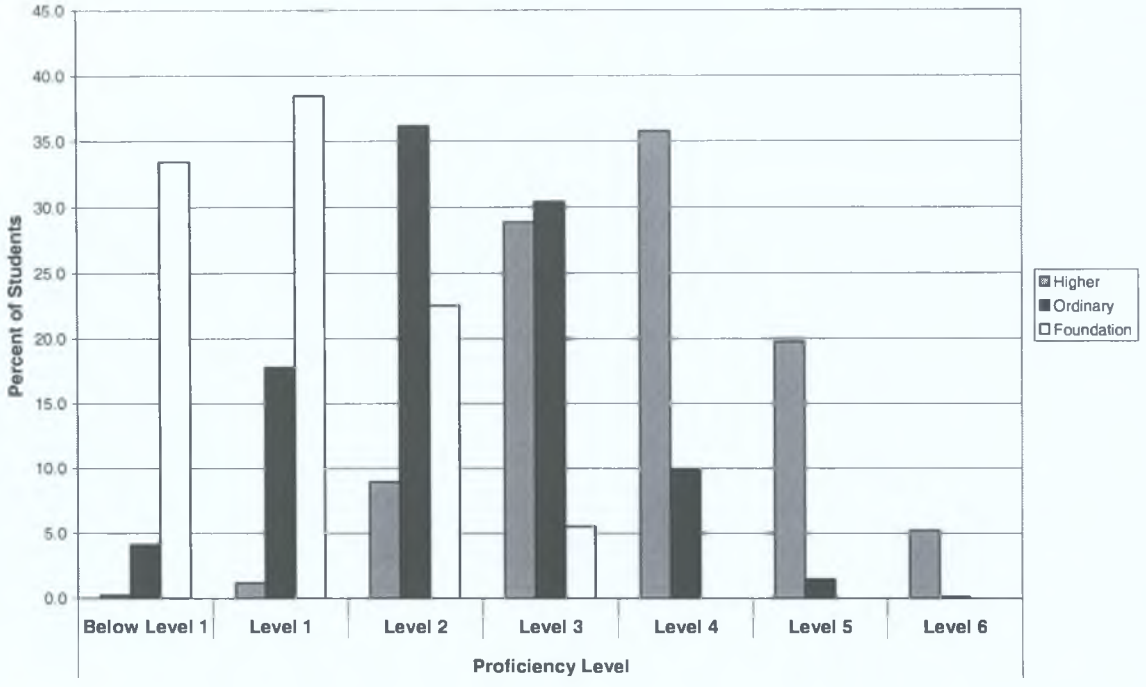


Table 2.11. Distribution of Students Taking Mathematics at Higher, Ordinary and Foundation Levels Across PISA Mathematical Proficiency Levels: PISA 2003

	Below Level 1		Level 1		Level 2		Level 3		Level 4		Level 5		Level 6	
	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE
Higher	0.3	0.16	1.2	0.33	9.0	1.09	28.8	1.26	35.8	1.52	19.7	1.39	5.2	0.76
Ordinary	4.1	0.70	17.8	1.26	36.2	1.27	30.4	1.30	9.9	0.90	1.5	0.38	0.1	0.11
Foundation	33.4	4.05	38.5	4.18	22.5	3.35	5.5	1.83	0.0	0.00	0.0	0.00	0.0	0.00

Source: Cosgrove et al., 2005, Table 6.19

2.3.3.2. Performance on Junior Certificate English and Mathematics at Higher, Ordinary and Foundation levels

Shiel et al. (2001) and Cosgrove et al. (2005) have reviewed the performance of Irish students on the Junior Certificate. They note that, in the 1999 English examination, 2.2% of students at higher level, 2.3% of ordinary level, and 8.6% of foundation-level students, were given a grade of E or F. In the 2003 English examination, the percentages are even lower, at 1.6%, 1.0% and 2.9%, respectively. In the 1999 Junior Certificate mathematics examination, 5.1% of higher-level students, 9.0% of ordinary-level students, and 7.7% of foundation-level students received grade E or F. In 2003, the corresponding percentages are 3.6%, 7.6% and 3.6%. Notwithstanding the different purposes of PISA test and the Junior Certificate Examination, these 'fail' rates are at odds with the percentages of students at or below Level 1 on the PISA reading and mathematics scales at ordinary and foundation levels.

At the other end of the scale, 31.8% of higher-level students, 32.2% of students at ordinary level, and 37.9% of students at foundation level received a grade A or B on the 1999 Junior Certificate English examination. The percentages for 2003 are 37.1%, 39.9% and 46.4%, respectively. In mathematics, the percentages awarded a grade A or B at higher, ordinary and foundation levels in 1999 are 44.0%, 36.0%, and 43.1%, respectively, and in 2003, 50.8%, 40.2%, and 53.2%, respectively. A comparison of the percentages of students attaining each letter grade with the percentages of students at each proficiency level suggests that the distributions of grades at higher level for both mathematics and English are more closely aligned to the distributions of students across the PISA proficiency levels.

2.3.3.3. Strength of the Association in Performance Between PISA and the Junior Certificate

In the PISA 2000 and 2003 national reports, Pearson correlations were reported for student performance on the Junior Certificate and on the corresponding PISA domain. These indicated highly consistent results across both domains and years. The Pearson correlations between PISA reading and Junior Certificate English in 2000 and 2003 were .74 and .67, respectively. The corresponding correlations for PISA mathematics and Junior Certificate mathematics were .73 and .75, respectively.

The strength of these correlations suggests that between 42% and 56% of the variance on the two achievement measures is shared, which suggests a moderate degree of overlap. However, correlations between the PISA domains of reading and mathematics are similar to these (.67 in 2000 and .80 in 2003), as are correlations between Junior Certificate English and mathematics (.71 in both 2000 and 2003[24]); clearly, other factors such as the style of the test, the item formats, testing conditions, the extent to which it is a high- or low-stakes test, are of potential relevance in considering these associations. Moreover, a Pearson correlation is an overall measure of association, which may not hold at the extremes of the achievement distributions.

[24] These correlations were not previously reported.

2.4. A Review of Between-Cluster Variance in Achievement in Ireland

Table 2.12 shows the percentage of total variance that is between schools/classes for various measures of achievement of students in Ireland, cited in 12 studies/surveys of both primary and post-primary schools published between 1976 and 2004 (with one in preparation). Each is reviewed in brief. The aim of the review is to ascertain whether, and to what extent, the interpretation of between-'school' variance should be considered with reference to the sample design (age- or grade-based), the subject area measured (school-dependent or school-independent), and/or the curriculum sensitivity and subject-dependence/independence of the achievement measure.

Madaus, Kellaghan and Rakow's (1976) study took as its starting point two problems with Coleman et al.'s (1966) early work on school effectiveness. First, that the measure used (generic verbal achievement) did not accurately reflect curricular/instructional objectives and hence was a poor measure of school effectiveness.[25] Second, the use of the school as a unit of analysis disguised differences within schools which may reflect differences in students' experiences, teacher differences, access to equipment, etc.

Madaus et al. (1976) compared, for a sample of boys in 32 post-primary schools, the between-school and between-class variance on a number of Leaving Certificate subjects, on an IQ measure, and on three standardised achievement measures. They found that comparatively small amounts of variance were between schools on the standardised measures and that, similarly, for the majority of Leaving Certificate subjects, between-school differences were not statistically significant (although the small number of schools in the sample may have contributed to the lack of statistical significance here). The strongest effects were associated with classes rather than schools, regardless of whether the test was curriculum sensitive (i.e. Leaving Certificate subjects) or not (i.e. standardised test measures). They interpreted this to indicate that ability streaming *within* schools has a greater impact on achievement differences than selectivity *between* schools.

[25] This argument should be interpreted in its wider context. Since the US does not have a national curriculum, standardised tests, which are as curriculum neutral as possible, had been developed, and used in the US for school effectiveness research and a variety of other purposes (Madaus, Airasian & Kellaghan, 1980).

Table 2.12. Summary of Studies Reporting Between-School/Class Variation in Achievement for Students in Ireland, 1976-2004

Survey/Year	*Sample*	*Achievement Measure*	*ICC*
Madaus, Kellaghan, & Rakow (1976)	32 post-primary schools, 49 classes, 1253 students (all male). Unit of analysis = class and school.	Primary Mental Abilities Test (standardised IQ test)	28.9 (school) 39.8 (class)
		Gates-McGintie Reading Test (standardised reading test)	19.9 (school) 48.2 (class)
		Graded Arithmetic Mathematics Test (standardised mathematics test)	27.8 (school) 49.9 (class)
		Leaving Cert. English (Higher)	21.2 (school) 43.6 (class)
		Leaving Cert. English (Lower)	27.3 (school) 26.5 (class)
		Leaving Cert. Irish (Higher)	73.3 (school) 7.7 (class)
		Leaving Cert. Irish (Lower)	29.3 (school) 25.1 (class)
		Leaving Cert. Mathematics (Higher)	44.7 (school) 28.6 (class)
		Leaving Cert. Mathematics (Lower)	33.6 (school) 37.2 (class)
		Leaving Cert. Mean	14.7 (school) 58.7 (class)
Kellaghan, Madaus, & Rakow (1979)	50 post-primary schools, 101 classes, 1560 students. Unit of analysis = class and school. Standardised test measures taken in First Year, Inter. Cert. in Third Year.	English 70 (standardised test of English	26.1 (school) 49.1 (class)
		Gaeilge 70 (standardised test of Irish)	29.8 (school) 51.5 (class)
		Graded Arithmetic Mathematics Test (standardised test of mathematics)	31.5 (school) 47.8 (class)
		Inter. Cert. English (Higher)	27.8 (school) 28.5 (class)
		Inter. Cert. English (Lower)	34.2 (school) 24.5 (class)
		Inter. Cert. Irish (Higher)	61.2 (school) 13.7 (class)
		Inter. Cert. Irish (Lower)	38.7 (school) 27.8 (class)
		Inter. Cert. Mathematics (Higher)	40.0 (school) 27.5 (class)
		Inter. Cert. Mathematics (Lower)	40.1 (school) 19.5 (class)
		Inter. Cert. Mean	31.6 (school) 34.9 (class)
Madaus et al. (1979)	50 post-primary schools, 101 classes, 1560 students. Unit of analysis = class. Standardised test measures taken in First Year, Inter. Cert. in Third Year.	English 70 (standardised test of English	43.4
		Gaeilge 70 (standardised test of Irish)	49.9
		Graded Arithmetic Mathematics Test (standardised test of mathematics)	47.0
		Inter. Cert. English (Higher)	26.3
		Inter. Cert. English (Lower)	29.9
		Inter. Cert. Irish (Higher)	44.8
		Inter. Cert. Irish (Lower)	36.8
		Inter. Cert. Mathematics (Higher)	38.4
		Inter. Cert. Mathematics (Lower)	26.2
		Inter. Cert. Mean	35.0

Table 2.12. Continued.

TIMSS (1995) (Beaton et al., 1996a, b; Foy, Rust, & Schleicher, 1996) Martin et al., 2000b)	132 post-primary schools; 129 first-year classes, 132 second-year classes; 3127 first-year students, 3076 second-year students. Classes identified on the basis of mathematics. Unit of analysis: most reports separated first and second years so, effectively, the class.	TIMSS Mathematics (international test for 42 countries that aimed to tap some common mathematics concepts)	52.0
		TIMSS Science (international test for 42 countries that aimed to tap some common science concepts)	38.0
Junior Certificate (1998) (Sofroniou et al., in prep.)	738 schools, ca. 63000 students who attempted the Junior Certificate in 1998. Unit of analysis: school.	Junior Cert. English	27.2
		Junior Cert. Mathematics	25.3
Smyth (1999)	116 post-primary schools; within each, 'base' classes selected (half of third year classes, half of sixth year classes); 5961 third years, 4813 sixth years. Unit of analysis: school.	Mean Junior Cert. score (on 10-point scale)	22.3
		Mean Leaving Cert. score (on 20-point scale)	19.8
PISA (2000) (OECD, 2001; Shiel et al., 2001)	139 post-primary schools; 3854 students aged 15 (born in 1984) sampled at random; 64.0% in Junior Cycle, 36.0% in Senior Cycle. Unit of analysis: school.	PISA Reading (international test using a literacy-based real-life approach)	17.8
		PISA Mathematics (international test using a literacy-based real-life approach)	11.4
		PISA Science (international test using a literacy-based real-life approach)	14.1
Junior Certificate (1999 and 2000) (Sofroniou, Shiel, & Cosgrove, 2000; Sofroniou, Cosgrove, & Shiel, 2002)	Students who participated in PISA 2000 and who attempted the Junior Cert. in 1999 or 2000 (94% of original sample)	Junior Cert. English	17.7
		Junior Cert. Mathematics	15.6
		Junior Cert. Science	16.2
PISA (2003) (OECD, 2005b; Cosgrove et al., 2005)	145 post-primary schools; 3880 students aged 15 (born in 1987) sampled at random; 63.7% in Junior Cycle, 36.3% in Senior Cycle. Unit of analysis: school.	PISA Reading	22.5
		PISA Mathematics	16.7
		PISA Science	16.2
National Assessment of English Reading (1998) (Cosgrove et al., 2000; Sofroniou et al., in prep.)	150 primary schools, 3886 pupils selected at random from fifth class. Unit of analysis: school.	Tasks for the Assesment of Reading Achievement (generic reading test including Narrative, Expository and Documents texts)	16.8
National Assessment of Mathematics Achievement (Shiel & Kelly, 2001; Sofroniou et al., in prep.)	120 primary schools, 4747 pupils selected from all fourth classes. Unit of analysis: school.	Test of Mathematics Achievement (based on 1999 primary mathematics curriculum; some links with TIMSS)	17.5
Survey of Disadvantaged Primary Schools (Eivers, Shiel, & Shortt, 2004)	94 designated disadvantaged schools with 2238 first clas pupils, 2120 third class pupils, and 2141 sixth class pupils. Unit of analysis: school.	The Drumcondra Sentence Reading Test (DSRT; a cloze-type multiple choice test of reading) - Level 1	18.1
		DSRT - Level 3	22.9
		DSRT - Level 6	13.5

One might add two further observations. First, between-school *and* between-class variance were both higher for higher-level Leaving Certificate subjects. This may relate to the fact that public examinations are designed to discriminate better between higher achievers compared with lower achievers (Greaney & Kellaghan, 1996; Martin & Hickey, 1992). Second, the fact that higher- and ordinary-level examination performance is not combined to a single scale makes comparisons between the standardised tests and Leaving Certificate subjects difficult.

In the second study shown in Table 2.12, Kellaghan, Madaus and Rakow (1979) compared the performance of students in 50 post-primary schools on several Intermediate Certificate (Inter. Cert.) subjects[26] and on three standardised test measures. The two issues with Coleman et al.'s work mentioned in the context of Madaus, Kellaghan and Rakow (1976) also guided the objectives of this study. At the end of the school year, students were administered standardised tests in English, Irish and mathematics (scaled to have a mean of 500 and standard deviation of 100). Two years later, examination results were available for 101 of the original 114 classrooms and 1560 of the original 2629 students. The analyses include only students for whom results are available for both sets of measures. Higher and ordinary syllabus levels were treated as separate groups; a mean Inter. Cert. score for each student was also computed.

Results indicated that variance between classes tended to be larger than variance within classes. Variance between schools was not negligible either. Kellaghan et al. (1979) interpret the results as an indication of the selectivity of the Irish post-primary school system, with selection occurring both between schools and between classes. They also interpret the tendency for between-school variance on the Inter Cert. subjects to be higher compared with the standardised test measures as an indicator of the differential effectiveness of schools in teaching material necessary to complete the Inter. Cert. Examinations. In contrast, the higher between-class variance associated with the standardised test measures is suggestive of ability streaming. They also note that higher between-school variance is associated with what they term school-dependent subjects (e.g., Irish, mathematics) compared with school-independent subjects (e.g., English).

[26] The Inter. Cert. was replaced by the Junior Cert. in 1989 (NCCA, 1989).

In the third study in Table 2.12 (Madaus, Kellaghan, Rakow, & King, 1979), which utilised the same sample and data as Kellaghan, Madaus and Rakow (1979) compared the between-class variance components associated with the Inter. Cert. Examinations and standardised test measures. It is important to note that Madaus et al. examined between-*classroom* rather than between-school variance. Their reasons for doing so have to do with the fact that between-school analyses assume that all students in a school are exposed to similar conditions, where it is more likely that students in different classes experience different teachers, curricula and physical resources.

However, there are also issues to consider when treating the classroom as the unit of analysis when examining achievement in post-primary schools in Ireland. (The general issues regarding the interpretation of achievements of a class-based sample compared with an age-based one were discussed in Chapter 1.) Smyth (1999) examined school effectiveness within a multilevel modelling framework and used the school as the unit of analysis, rather than the class. She does acknowledge that a three-level model (students in classes in schools) would have been 'illuminating', but questioned its feasibility. The Junior Cycle students in her survey were in the same class for three or fewer subjects in one-third of the schools. She argues that "many pupils in second-level schools have no 'class' in any real sense" (p. 22). This argument does not hold to the same degree, however, when one is considering performance in a specific subject, such as mathematics, where the concept of class *is* meaningful, provided the sample design entails selection of pupils by intact class pertaining to the subject in question. A second reason given by Smyth for choosing the school over the class is that the models are retrospective: students in a school have been exposed to different teachers and classmates, which is in part dictated by school-level policy on subject availability, streaming, etc. Hence it is reasonable to say that both approaches have their merits and complications, and the decision as to which approach to use should be made with reference to the sample design and the aims of the study/analyses. If intact class sampling with at least two classes per school is used, either the school or the class (or both) may be used as the unit of analysis; if only one class per school is used, then the interpretation of the between-cluster variance is by default that which is between classes, and if random within-school sampling is used whereby students are sampled across multiple classrooms, then the default unit of analysis is the school.

It is unfortunate that three-level statistical modelling, which allow the partitioning of variance into three levels (in this instance, student, class and school) was not in use in educational research at the time of Madaus et al.'s study[27] since the technique may have allowed the simultaneous examination of school, class and student effects, particularly given that more than one class was selected in each school. In fact, of the 16 classroom variables examined in their study, three are school-level variables (percentage of boys in the school, school size, number of examination subjects offered by the school), and a further two are based on principals' opinions rather than those of teachers; and of the five individual/classroom variables, two could be considered school-level variables in the context of a three-level model (attendance in a secondary school; attendance in a vocational school) (although of course they may also reflect parental or student preference as well as being a characteristic of schools).

In any case, Madaus et al. (1979) found, contrary to their hypothesis, that between-class variance for the standardised measures of English, Irish and mathematics was quite high (ranging from 43% to 50%); the between-class variance for the Inter. Cert. subjects tended to be lower (ranging from 26% to 45%). However, there was a second aspect to this study, which entailed a comparison of the explanatory power of school- and class-level factors of between-class variance (described in Section 2.5). Since the standardised measures were taken within one year of student intake into post-primary, Madaus et al. postulated that this pattern of results may reflect selectivity factors rather than (or in addition to) instructional/curricular ones. A second observation may also be made about these data, which is that, consistent with Madaus et al. (1976), between-class variance was higher for the higher-level Inter. Cert. subjects (English, Irish, mathematics). The between-class variance for higher-level Irish (about 45%) is higher than for both English (26%) and mathematics (38%). Madaus et al. suggest that the lower-level subjects might be more subject to general than specific influences. To this one might add the likelihood that (i) as noted, the Certificate Examinations discriminate better between candidates with high levels of performance than those with lower levels, and hence variance in higher level subjects are more amenable to explanation; and (ii) that separating higher and lower levels of the subject out might confound the classrooms

[27] Snijders and Bosker (1999, pp. 1-2) suggest that multilevel models were not generally in use until after 1980, and that the basis of multilevel analysis was not established until 1986.

students are in, in schools where students are streamed into different classes according to syllabus level.

Turning now to the fourth study in Table 2.12, the TIMSS 1995 sample comprised all students enrolled in the two adjacent grades that contain the highest proportion of 13-year olds. Within schools, one intact class was sampled at each of the grade levels (in Ireland, these were first and second year). Foy, Rust and Schleicher (1996, p. 4-7), note that ideally, students should be in the same class for both mathematics and science, but in practice this was not the case in several countries including Ireland. In such cases, classes were identified on the basis of mathematics.

The between-school variance for mathematics in Ireland was 52%; for science, 38% (Martin et al., 2000b). Out of 34 countries compared, Ireland had the ninth highest between-school variance for mathematics, and the seventh highest for science. Since the sample was of intact mathematics classes, it is not surprising that the between-school variance is higher for this subject. These figures are also broadly comparable with figures for the standardised tests used in Madaus et al. (1979). However, it is not made clear in Martin et al.'s discussion of the results (pp. 74-75) that they are in fact speaking about between-*class* variance, in the sense that Madaus et al. (1979) were, since the figures reported by Martin et al. are based on the sub-sample of students in the higher grade level (grade 8 or second year) rather than all participating students. I will return to this issue in Chapter 5, in a re-analysis of the TIMSS data which includes a comparison with the mathematics achievement data from PISA 2000.

Sofroniou et al. (in preparation) reported the percentage of variance in achievement that is between schools for *all* students taking Junior Certificate English and mathematics (scaled to a 12-point Junior Certificate Performance Scale; JCPS) in 1998 as 27.2% and 25.3%, respectively. These are slightly lower than Madaus et al.'s (1979) study would have suggested, but the replacement of the Intermediate Certificate with the Junior Certificate in 1989 (NCCA, 1989), along with the placement of the Junior Certificate outcomes on one rather than two scales, makes direct comparisons between these two studies difficult if not impossible.

The results from Smyth's (1999) study are closer to those reported in Sofroniou et al. (in preparation). In her study of school effectiveness in Irish post-primary schools, Smyth sampled 116 schools and, within each school, roughly half of the intact classes from third year (Junior Certificate year) and sixth year (Leaving Certificate year). The average of examination subjects taken by students was used as the outcome measure, on a 10-point scale for the Junior Certificate, and a 20-point scale for the Leaving Certificate. The percentage of total variance that was between schools the Junior Certificate was 22.3%; for the Leaving Certificate, it was 19.8%. The latter figure is comparable to if a little higher than the figure reported by Madaus et al. (1976) for the mean Leaving Certificate performance (14.7%), which is perhaps related to the increase in retention rates at upper post-primary level. For example, during the 1979-1980 school year, 68.0% of 16-year-olds, 49.6% of 17-year-olds, and 25.9% of 18-year-olds were in full-time education in Ireland. In the 1999-2000 school year, the corresponding figures are 91.0%, 81.2%, and 61.8%, respectively (Ireland, 1981; 2001).

In PISA 2000, the between-school variance in Ireland for all three domains assessed was low relative to the other countries surveyed: for reading, it was 17.8% (OECD average = 34.7%), for mathematics, it was 11.4% (OECD average = 31.4%), and for science, it was 14.1% (OECD average = 30.6%). The between-school variance for PISA 2003 was also comparatively low: 16.7% for mathematics (OECD average = 32.7%), 22.5% for reading (OECD average = 31.4%), and 16.2% for science (OECD average = 29.9%) (Shiel et al., 2001). In an analysis of performance on Junior Certificate English, mathematics and science of students who participated in PISA 2000 and who took the examination in either 1999 or 2000, the between-school variance was similarly low to the PISA measures: 17.7% for English, 15.6% for mathematics, and 16.2% for science (Sofroniou, Shiel & Cosgrove, 2000; Sofroniou, Cosgrove & Shiel, 2002). The exception is the figure for Junior Certificate mathematics (15.6%), which is a little higher than that for PISA mathematics (11.4%). The figures for the Junior Certificate are perhaps lower than expected, given that curriculum-sensitive measures might be expected to be more prone to between-school variance, and lower than the results reported previously by Kellaghan et al. (1979). This suggests that the age-based sample may depress between-school variance on curriculum-sensitive measures.

Although not of central relevance to the present study, it might also be noted that between-school variance in achievement is also quite low at primary level (Sofroniou et al., in preparation). The percentage of variance on a generic measure of reading skills in the 1998 reading survey, which involved a random sample of 5th class pupils in 150 schools, was 16.8% (Cosgrove, Kellaghan, Forde & Morgan, 2000). In the 1999 mathematics survey, in which assessment is largely based on the 1999 primary mathematics curriculum and involving 120 schools and a sample of all 4th class pupils within each school, it was 17.5% (Shiel & Kelly, 2001). In a survey of literacy in disadvantaged schools, the between-school variance in a sentence reading test was 18.1% at first class level, 22.9% at third class, and 13.5% at sixth class (Eivers, Shiel & Shortt, 2004).

This review of the percentage of total variance that is between schools/classes in Ireland is complicated by virtue of the number of years which the studies span and the fact that, in earlier studies, achievements of higher- and ordinary-level Inter. Cert. and Leaving Certificate candidates were not combined. However, some general observations may be made. First, variance between schools (as opposed to classes) is, with the exception of higher-level Irish at both Inter. and Leaving Cert. levels, below 45%, regardless of the subject area, and also regardless of whether the achievement measure is intended to be curriculum sensitive or not. Second, a comparison of between-school (as opposed to class) variance associated with surveys of post-primary schools conducted in the 1970s compared with those in the 1990s and later suggest an overall drop in between-school variance in achievement, particularly in mathematics. This can be attributed in part at least to the increased retention rates at both lower and upper post-primary levels in Ireland over the past 30 years, although curricular changes also make these comparisons rather complex. Third, between-class variance is generally substantial, and comparable to, if not greater than, between-school variance. This is strongly indicative of within-school selection by ability or other related characteristics. However, there is considerable variability in the absolute values associated with between-school and between-class variance which suggest that attention should be paid to the subject area and curriculum sensitivity of the measure. It appears to be the case that standardised test measures are associated with higher between-class variance but lower between-school variance (which may be explained by class allocation based on student ability), while, school-dependent and curriculum-sensitive measures are associated with both higher

between-class and between-school variance (which may be explained by the higher sensitivity of school-dependent subjects to school and class effects). Fourth, in international terms (i.e., comparing PISA 2000 and 2003 with TIMSS 1995), it would appear that between-school variance in post-primary schools in Ireland is comparatively low, whereas between-class variance is comparatively high. This suggests that estimates of variance components based on grade-based samples do indeed confound between-school and between-class variance to a substantial degree in Ireland, and that estimates of variance components based on an age-based sample may disguise substantial achievement variance that is between classrooms within schools. However, differences in the content of the tests and in the target populations of PISA and TIMSS should be noted in making these inferences. Fifth, it is perhaps not surprising that variance components for PISA 2000 reading and Junior Certificate English for students participating in PISA 2000 are similar, given that the manner in which PISA assesses reading is congruent in many respects to Junior Certificate English, and that reading is a more generic, less school-dependent measure in any case (particularly at the age of 15). Perhaps more surprising is the finding that the variance components for Junior Certificate mathematics for students participating in PISA 2003 are only a little higher than those for PISA 2003 mathematics, given the large divergences between the content and style of the two assessments described earlier in this chapter (that is, one would have expected the between-school variance for Junior Certificate mathematics to be considerably higher than for PISA mathematics). One possible explanation for this smaller-than-expected difference is that the age-based sample design disguises between-class differences in the achievement measures; a much bigger difference might have been observed, had PISA employed a grade-based design and the selection of intact classes.

2.5. A Review of Explanatory Statistical Models of Student Achievement in Ireland

In this section, I review a study by Madaus et al. (1979) which compares between-school/class variance in achievement on a variety achievement measures. I then review some recent 'explanatory' models of achievement in Ireland from international surveys (TIMSS 1995, PISA 2000, PISA 2003), secondary analyses of the PISA datasets involving students' Junior Certificate Examination performance, and national surveys (Sofroniou et al., in preparation; Smyth, 1999). All studies described here, with the exception of Madaus et al.'s, use multilevel modelling. Notwithstanding the general

limitations associated with these models noted in Chapter 1, the primary aim of this section is to investigate whether there is evidence of differences in the amount of explained variance associated with class/school-level variables (after adjusting for student social background and school social intake) which may be related to the subject area, the extent to which the achievement measure is aligned to the curriculum, and whether the sample design is grade-based or age-based, since there is no research in the past 25 years which directly addresses these issues. The review also considers whether the impact of social intake varies according to the sample design and achievement measure.

As noted previously, Madaus et al. compared performance on standardised and curriculum-based (public examination) measures of educational achievement and hypothesised that between-school variance would be higher for the curriculum-based assessments. They also hypothesised that school-level (or class-level) variables would explain more of the variance in achievement on the curriculum-based measures than on the standardised tests, since the former could be expected to be linked more closely to instructional practices and other characteristics of classes and schools.

Madaus et al. divided predictor variables into five blocks – individual, classroom, individual/classroom, family background, and IQ (i.e., a control for student general scholastic ability), and identified 42 out of the original 82 using a strategy which eliminated non-significant variables from each block. There are some difficulties with this approach within an explanatory statistical framework, however, since this method, a commonality analysis, although commonly used in the 1970s (e.g., Purves, 1975, in analyses of data from IEA studies; Mayeske et al., 1969, in a re-analysis of the Coleman et al., 1966 data), is not useful for explaining achievement (being better suited to predictive models). Further, stepwise regression favours prediction with relatively large unique contributions, resulting in blocks made up of homogenous predictors, and runs the risk of excluding variables that may be substantively significant.

The classroom block generally explained a considerably larger proportion of the unique variance on the curriculum-based measures than on the standardised test measures. It comprised percentage of boys in the school, school size, mean time spent on homework, number of examination subjects offered by the school, teacher educational expectations,

student participation in school activities, students' perceptions of teachers; expectations regarding conformity to academic press and discipline, percentage of students that the principal would feel 'better off' in another school, principal opinion on optimal school size, and availability of counselling service. This finding holds both for total variance and for between-class variance (Table 2.13), and supports their second hypothesis. They noted that Inter. Cert. Irish seemed particularly sensitive to classroom influences. In passing, it might also be noted that mathematics was also relatively sensitive to such effects, and more so than English, which is consistent with the argument that English is less school-dependent than mathematics.

Table 2.13. Between-Class Variance for Standardised and Curriculum-Sensitive Tests, Total and Between Class Variance, and Percent of Explained Variance Attributable to Class-Level and Family Factors (Madaus et al., 1979)

Achievement measure	*% of total variance between classes*	*% of total variance explained (all 5 blocks)*	*% of total variance explained uniquely by class factors*	*% of total variance explained uniquely by family factors*	*% of between-class variance explained (all 5 blocks)*	*% of between-class variance explained uniquely by class factors*
Standardised English	43.4	65.4	4.0	1.2	92.9	6.0
Standardised Irish	49.9	50.4	17.0	1.2	81.3	9.2
Standardised Mathematics	47.0	53.5	8.0	0.7	84.0	17.4
Inter. Cert. English (Higher)	26.3	21.9	40.6	--	63.4	33.8
Inter. Cert. English (Lower)	29.9	25.6	21.1	4.7	39.5	18.1
Inter. Cert. Irish (Higher)	44.8	43.9	67.9	4.3	82.6	66.5
Inter. Cert. Irish (Lower)	36.8	31.1	34.7	10.3	51.2	29.3
Inter. Cert. Maths (Higher)	38.4	50.3	60.2	9.9	98.7	78.9
Inter. Cert. Maths (Lower)	26.2	30.8	24.0	1.9	60.3	28.2
Inter. Cert. Mean	35.0	52.2	22.2	0.8	91.7	33.1

Source: Madaus et al., 1979, Tables 1, 3 and 4.

Variables relating to family background (family size, position in family, father's occupation on a five-point scale, various parental educational expectations) generally contributed little to the explained variance (see again Table 2.13), but Madaus et al. noted that some measures used in other studies, such as parental education and books in the home, were not included in their measures. Nonetheless, this finding is at odds with more recent studies reviewed later in this section, which found strong effects associated with social background.

I turn now to the more recent multilevel models of achievement, beginning with TIMSS 1995. One of the TIMSS 1995 reports concerned school effects (Martin et al., 2000b)

and included multilevel models of the mathematics and science achievements of grade 8/second year students. The models indicate that a substantial amount of the variance between schools in Ireland is explained by home background (number of people living at home, number of parents at home, books in the home, material and educational possessions, and highest level of education of mother and father) – 51% in the case of mathematics, and 52% in science. Martin et al. also reported the results of a series of explanatory analyses relating to five groups of variables: classroom practices, teacher characteristics, school climate, school location and size, and home-school interface. It should be noted that the classroom block included self-ratings of attitudes to mathematics and science. The circular nature of the relationship of these types of variables with achievement has been noted (Cosgrove et al., 2005). The results of the models for Ireland indicate that classroom characteristics explain 67% of between-school variance in mathematics achievement and 61% in science achievement; that the addition of variables relating to teacher characteristics, school climate, and school location and size do not explain any additional variance in either model; that the addition of home-school interface variables explains an additional 9% of between-school variance in mathematics and 6% in science; and that the addition of home background to the models of mathematics and science explains an additional 4-5% of between-school variance. A comparison of models with and without home background suggests that, over and above home background, the other variables explain an additional 28% of variance in mathematics, and 23% of variance in science. This suggests substantial covariation between social background and school/class variables. It also suggests that mathematics is somewhat more sensitive to school/class effects than science, but this may be confounded with the sample design which entailed intact class sampling based on mathematics class.

In the initial international PISA 2000 report (OECD, 2001b), three three-level models (one for each domain assessed) for all countries combined were presented. Each included four blocks of variables (family background and student characteristics; school resources; school policy and practice; and classroom practice). Altogether, for the model of reading, these explained 43.4% of total variance in achievement between countries, 71.9% of variance between schools, and 12.4% of variance between students. Social background explained substantial percentages of the three components (34.3%, 66.1%, and 12.4%, respectively) and in fact the addition of the school and class

variables added no additional explained variance at the student level. These models, unfortunately, tell individual countries nothing about variance in achievement between schools and students; they are also problematic in the sense that it is assumed that countries are exchangeable entities, and that the variables in the model are comparable in meaning and relevance for each country. The model also neglects the possibility that other relevant country-specific variables have been excluded.

In one of the thematic reports on PISA 2000, school factors associated with quality and equity were analysed (OECD, 2005c). The report included a number of multilevel models carried out on a country-by-country basis which examined the relative impacts of student characteristics, school context, school climate, school resources and school policies on reading achievement. Student characteristics examined were SES, gender, age, immigration status, grade level, and study programme. School context was measured by type (public or private), location, and average SES. School resources comprised quality of the school's building, educational resources, computer resources, teacher qualifications, perceived teacher shortage, student-teacher ratio, and teacher professional development. School climate measures were disciplinary climate, teacher support, achievement press, student-teacher relations, students' sense of belonging in school, student behaviour, and teacher morale. School policies examined were instructional time, policies on student progress, self-evaluation, transfer, admission and placement policies, policy on the use of performance information, school autonomy, and teacher autonomy. In the model for Ireland, student characteristics explained 40% of between-school variance, school social intake 33%, and the remainder (school climate, policies and resources) just 2% of variance (which was not statistically significant). The corresponding OECD averages were 50%, 24%, and 8%, respectively (OECD, 2005c, Table 3.2, p. 119). Thus in Ireland, 'policy amenable' school factors as measured in PISA appear to have little explanatory power. There are two limitations with these analyses as they apply in the Irish context. First, some of the variables are not particularly relevant in Ireland (e.g., instructional time which is standardised). Second, variables relating to national structural features of the educational system (e.g., school sector) are not included in the models.

In the PISA 2003 international report (OECD, 2004c), information was provided for each country on the amount of variance in achievement in mathematics between and

within schools that is explained by student and school economic, social and cultural status (ESCS), by student study programme, single parent status, country of birth, and language spoken. (The results of these models were already discussed for illustrative purposes in Chapter 1.)

Because of the limitations associated with the international PISA models, the national reports for PISA 2000 (Shiel et al., 2001) and PISA 2003 (Cosgrove et al., 2005) included multilevel models of the Irish data which incorporated a wider range of nationally-relevant variables. Since the two domains of interest are reading and mathematics, models for science are not reviewed here.

In PISA 2000, the final model for reading included student gender, SES (i.e., parental occupation, which had a weak curvilinear association with achievement), number of siblings, books in the home, absence from school, completion of homework on time, leisure reading, attitude to reading, and dropout intent (the effect of which varied across schools); and at the level of the school, disciplinary climate, school sector, and designated disadvantaged status. The model included an interaction term for books in the home and gender (where the gender difference, favouring females, is greater for students with higher amounts of books in the home). A number of variables were dropped from the model in the course of model building (parental education, diversity of reading, lone parent status, parental engagement, school gender composition, school size, and student-teacher ratio). The final model explained 77.8% of variance between schools, and 44.2% within schools.

The model for PISA 2000 mathematics explained 78.8% of variance between schools, and 31.9% of the variance within schools. The variables in the final model were student gender, SES, parental education, lone parent status, number of siblings, books in the home, dropout intent, completion of homework on time, and grade level; and at the school level, school sector and disciplinary climate. The model also included an interaction term for lone parent status and student gender (where a larger gender difference, favouring males, was associated with students from lone-parent households).

There are considerable commonalities across the two national models for PISA 2000. First, the proportion of explained variance is similar, whereby the models explain the

majority of between-school variance but less than half of the variance within schools. Second, several variables relating to SES, family structure, and home educational environment appear in the final models. Third, school sector and designated disadvantaged status appear in all three models, explaining significant variance in achievement over and above student social background. This is suggestive of differential school effects.

In the PISA 2003 national report (Cosgrove et al., 2005), a similar list of candidate variables was selected for inclusion in the models. One important difference, however, is that, rather than using the binary measure of school designated disadvantaged status, the percentage of students entitled to a fee waiver for the Junior Certificate Examination, which is a proxy for school-level economic deprivation, was used. Furthermore, in addition to testing for interactions between student gender and the other student-level variables, all two-way and cross-level interactions were examined. The report also provides the additional explained variance due to the school-level variables (this information was not provided in 2000).

The final model of PISA 2003 mathematics included the following student-level variables: gender (males outperformed females), SES, lone parent status, number of siblings, number of books in the home (which interacted with absence from school), home educational resources, absence from school, and grade level; and the following school-level variables: disciplinary climate and fee waiver. School sector did not appear in the final model. The two school-level variables only explained an additional 16.5% of school variance and 2.7% of variance at the student level. The variance explained by the final model was 78.8% between schools and 29.6% between students. The same variables appeared in the final model of PISA 2003 reading, which did not require any interaction terms; the only difference of note is the direction of the gender difference, which favours females rather than males. The addition of the school-level variables explained an additional 20.1% of between-school variance and 4.2% of the variance within schools.

The national analyses of the PISA data, unfortunately, do not give an indication of the amount of variance in achievement that is uniquely explained by student and school social background; nor do they indicate the degree to which social background and

other factors covary, so one cannot determine the substantive significance of variables in the models. Further, the paucity of variables measuring Type B effects limits the extent to which the models provide insights into differential school effects relating to school and class variables. Using the same set of explanatory variables across each subject domain also reduces the chances of finding differences had a more iterative approach been used.

In secondary analyses of the PISA 2000 data, Sofroniou et al. (2000; 2002) used the same set variables from the PISA 2000 dataset to model the Junior Certificate achievements of Irish students in English, mathematics and science as was used with the models of achievement on PISA. The Junior Certificate grades were put on the 12-point Junior Certificate Performance Scale (JCPS).

The model for Junior Certificate English explained 79.3% of between-school variance and 37.3% of within-school variance. Over and above the student-level variables in the final model, the school-level variables explained an additional 17.3% of between-school variance, and 3.2% of variance within schools. The set of variables in the final model is very similar to the model for PISA 2000 reading, as is the magnitude of the parameter estimates associated with them, except that lone parent status and parental engagement remained in the model of Junior Certificate English, and the model required an interaction term for attitude to reading and gender in addition to books in the home and gender. These slight differences suggest that home background variables may be more relevant in interpreting achievement differences in Junior Certificate English compared to PISA reading.

The final model of achievement on Junior Certificate mathematics explained 64.3% of variance between schools and 29.5% within schools; the school-level variables explained an additional 10.2% of achievement variance at the school level and just 1.7% at the student level. The variables in the final model of Junior Certificate mathematics are similar to those in the final model for PISA 2000 mathematics, but without number of siblings and an interaction between lone parent and gender, and with the addition of parental engagement, absence from school, and an interaction between gender and completion of homework (where males outperform females, with a larger gender difference associated with lower frequencies of homework completion). That two

variables relating to engagement in school (i.e., absenteeism and completion of homework) appear in the final model of Junior Certificate mathematics but not in PISA mathematics is noteworthy and suggests that achievement on Junior Certificate mathematics, but not on PISA mathematics, is sensitive to school-related behaviours of students. When tested alone, the gender difference was not significant, which contrasts with PISA mathematics, where males significantly outperformed females.

A comparison the models of achievement on PISA 2000 and achievement on Junior Certificate English and mathematics indicates that the models explain similar proportions of achievement variance between and within schools, that the gender differences are consistent in the case of English/reading but not in the case of mathematics, and that the variables retained in the models are broadly similar. However, the same set of candidate explanatory variables (which included relatively few school-level variables) was used in these analyses. This constrains the extent to which differences may have been found. None of the models indicate the *relative* proportions of variance explained by SES and other variables, and there may have been some differences.

Sofroniou, Shiel and Cosgrove (2002) expanded the model of PISA 2000 reading to include seven attributes of students' self-regulated learning. These are a series of composite variables based on students' self-ratings. Only four of these remained in the final model, however (competitive learning, co-operative learning, instrumental motivation, and academic self-concept), and the relative contributions of these variables to student scores was quite small. The analysis serves as a critique of the prominence given to the self-regulated learning variables in the international report for PISA 2000, and the publication of a thematic report devoted to the self-regulated learning variables (OECD, 2001b; 2003a). Elsewhere (Cosgrove et al., 2005) the inclusion of these types of variables in explanatory models has been criticised, particularly those entailing a self-rating of efficacy or confidence in performance in a subject domain, since it is likely that their relationship with achievement is circular.

I turn now to a set of models developed in a national study which aimed at refining procedures for selecting schools for targeted intervention against educational disadvantage. Sofroniou et al. (in preparation) presented the results of four multilevel

models which were designed to address questions about the strength and nature of the social context effect in Irish schools. This was investigated to see "whether the relationship between levels of concentration of disadvantage in a school was linear or whether there were non-linear discontinuous relationships that could be taken to suggest meaningful thresholds for making decisions about targeted resource allocation" (p. 20). Failure to find a context effect, on the other hand, would favour policies targeting individual children rather than schools. They present results of two models at primary level and two at post-primary level. The primary level models used data from the 1998 National Assessment of English reading of fifth class pupils and the 1999 National Assessment of Mathematics Achievement of fourth class pupils. The post-primary level models used data from the 1998 Junior Certificate Examinations database; specifically, 12-point JCPS scores for English and mathematics. The models also included student gender, whether the student's family was in receipt of a medical card, and the percent of medical card holders in the school.

Results indicated firstly, that student gender and medical card status explained up to 32% of variance between schools. The percentage of explained variance tended to be slightly lower for mathematics at both primary and post-primary levels. Second, the percentage of medical card holders in schools (school social intake) explained up to an additional 29% of the variance between schools. In total, the variables explained up to 59% of between-school variance. Less explained variance was associated with the mathematics measures. Third, the models only explained a small proportion of within-school variance – between 11% and 22%. Explained within-school variance was higher for the post-primary models and slightly lower for mathematics compared with English/reading. Fourth, in all of the models, the social context effect was linear. Fifth, some cross-level interactions were found. For fourth class mathematics, the slope is steeper for males; for fifth class reading, the slope is steeper for non-medical card holders; for Junior Certificate mathematics, the slope is steeper for males and non-medical card holders; and for Junior Certificate English, there is again a steeper slope for males. Taken together, the four models show that the social context effect affects both medical card holders and non-holders, as well as males and females. Contrary to arguments put forward by Willms (2002), which suggest a stronger context effect for 'minority' or lower-status groups, the models reported in Sofroniou et al. suggest that the social context effect may operate more strongly on non-medical card holders.

However, consistent with Willms (2002), the social context effect in three of the four models is stronger for males. Sofroniou et al. acknowledge that multiple indicators of SES may have been preferable, but cite the strong correlations between indicators as a basis for using the single indicator. They also point out that other indicators do not have the same meaning in urban and rural settings.

The final set of multilevel models reviewed here is taken from Smyth's (1999) study of school effects in Irish post-primary schools. She examined various factors that accounted for achievement on the Junior Certificate (average Junior Certificate score on a 10-point scale) and the Leaving Certificate (average Leaving Certificate score on a 20-point scale).[28] Using an average score on these examinations means that the models cannot allow us to draw any conclusions about the differential results that might have been observed had subject areas been treated as individual outcomes. Smyth (1999) incorporated a measure of ability as an adjustment for 'intake' (scores on a verbal reasoning test), but the ability measure was taken just three months before the Junior Certificate Examination and therefore is probably confounded with school effects. She examined the percentage of achievement variance explained by various combinations of factors.

Using average Junior Certificate performance as the outcome, student background (gender, social class on a five-point scale, parental education and an indicator of whether pupils are older than the modal age) explained 52.6% of variance between schools and 15.0% of variance within schools. Together, pupil background, the ability measure, and school social context (the average of the five-point social class scale) explained 81.9% of variance between schools and 58.7% within schools. Adding school type (sector) to the model explained less than 1% of additional variance at school and student levels and is not significant. However, the addition of a block of variables relating to school organisation and processes, or school practices explained an additional 4.1% of between-school variance and 8.3% of within-school variance. The variables that were significant in this block were membership of top, middle or bottom stream class, pupil behaviour, teacher-pupil interaction, teacher expectations, and pupil aspirations.

[28] She examined other outcomes including absenteeism and stress levels, but the achievement outcomes only are reviewed here.

A similar approach was used to analyse achievement on the Leaving Certificate, but this time, the Leaving Certificate results were modelled using both Junior Certificate scores and verbal ability scores. Pupil background explained 41.6% of variance between schools and 6.5% of variance within schools. Pupil background taken with the two prior achievement measures explained 80.8% of variance between schools and 64.0% within schools. Social context explained just 3.8% of additional between-school variance and no additional within-school variance. Adding school type (sector) to the model explained just 0.1% of additional variance at school and student levels and was not significant. The addition of a block of variables relating to school organisation and processes explained an additional 5.1% of between-school variance and 4.3% of within-school variance.

Smyth (1999) comments that most of the achievement differences observed are accounted for by intake and social background, and her technique of partitioning variance by addition of separate blocks of models is useful in determining the magnitude of the Type B effect. Differences between school types are accounted for by differences in student background and school social context at both Junior and Leaving Certificate. The models for Junior and Leaving Certificate are similar. The school social context effect is smaller than Sofroniou et al.'s models might suggest once one factors in student ability as an intake measure, but since Smyth's ability measure is contemporaneous with her outcome measure for the Junior Certificate, it is likely that the intake measure results in somewhat of an underestimate of social context effects.

To summarise, the explanatory models reviewed here, with the exception of Madaus et al.'s (1979), were not explicitly designed to investigate whether the nature of the outcome measure and sample design produce different results. The results of the earlier models may not still hold due to substantial changes in the Irish education system in the past 30 years (e.g., curricular revisions, retention rates), and the treatment of higher- and ordinary-level groups differs to more recent models of the public examinations, which use various methods to combine the examination results of different syllabus levels into a single group. Further, the partitioning of variance components of these earlier studies on the basis of classes is not comparable with the PISA survey which uses age-based samples. Notwithstanding these difficulties, the earlier studies provide some evidence of

differential school and class effects on standardised and curriculum-sensitive measures. They also suggest that school-dependent subjects are more sensitive to school- and class-based effects than school-independent subjects (although the potential confounding of syllabus level with class membership should be noted). Common themes underpinning many of the explanatory analyses reviewed here are (i) the extent to which social background explains achievement (ii) after adjusting for social intake, the extent to which school/class practices explain achievement. Despite the research evidence from Madaus et al. which suggests that both the achievement measure and the sample design are relevant in considering these results, however, the OECD reports make no reference to these characteristics of the survey, and it may be the case that different conclusions may be drawn on these two themes, had the test and/or sample design of PISA been different.

2.6. How the Proposed Analyses Add to Existing Research

This section considers how the analyses in Chapters 3, 4 and 5 address some of the major issues raised in Chapters 1 and 2. Table 2.14 shows the three broad characteristics of the PISA survey which are considered in this study. These aspects will be used to address three questions:

- What does PISA tell use about the *achievements* of students in Ireland?
- What does PISA tell us about the *equity* of achievement outcomes in Ireland?
- What does PISA tell us about the *determinants* of achievement in Ireland?

The remainder of this section explains how the themes described in Table 2.14 address these questions.

2.6.1. What Does PISA Tell us About the Achievements of Students in Ireland?

Two aspects of this issue are considered. The first is a consideration of the extent to which the PISA measures of achievement differ to national achievement measures (the Junior Certificate Examinations). Much of the material for drawing conclusions about this theme comprises existing research, reviewed in this chapter. It has been shown that, in broad terms, the PISA reading test is in accordance with the Junior Certificate English syllabus and so may be used in a relatively straightforward manner to draw conclusions about the effectiveness of the Irish education system and to interpret Irish student performance in terms of international benchmarks. In contrast, PISA mathematics diverges considerably from Junior Certificate mathematics, and as such,

problems present in drawing conclusions about the relative effectiveness of the system; i.e., what does it mean to be at the OECD average on an assessment of mathematics which differs in many ways to what is assessed in the Junior Certificate Examination? This is not to say that the PISA mathematics achievement outcomes are not desirable; rather it is to say that differences between it and Junior Certificate mathematics need to be considered when using the results to develop policy on mathematics education in Ireland. Thus, the interpretation of Irish student performance in terms of international benchmarks is substantially more complex for mathematics than for reading. One way of examining this is to consider the extent to which achievements on PISA and the Junior Certificate are related to one another. This has been done for both PISA and 2003, but it was shown that intra-assessment correlations (i.e., PISA reading with PISA mathematics, Junior Certificate English with Junior Certificate mathematics) are about the same as correlations between assessments (i.e., PISA reading with Junior Certificate English; PISA mathematics with Junior Certificate mathematics). Moreover, correlations only provide an overall measure of association; the relationship may not hold at the extremes of the achievement distribution.

There are some limitations to the existing 12-point Junior Certificate Performance scale (JCPS) which has been used in analyses of links between the performance of students on PISA and the Junior Certificate. These limitations are explained here and it will be shown how the analyses in Chapter 4 attempt to address them.

Whether the JCPS used in many analyses of Junior Certificate achievements are appropriate or not have not been thoroughly investigated, nor has any research been conducted which investigates what the linkages may mean. A more thorough analysis of this issue may shed more light on interpreting the PISA benchmarks with respect to national achievements as measured on the Junior Certificate Examinations, which is of interest given the lack of national assessments at post-primary level. The 12-point scale which has been used in analyses of the Junior Certificate in both PISA 2000 and PISA 2003 (Cosgrove et al., 2005; Shiel et al., 2001) is based on work by Martin and Hickey (1992; 1993) in reviews of performance students on the 1992 Junior Certificate and 1991 Leaving Certificate Examinations. Kellaghan and Dwan (1995) also used the 12-point scale in analyses of results of the Junior Certificate, while Millar and Kelly (1999) used the scale in a longitudinal study of students taking the Junior Certificate

Examination in 1994 and the Leaving Certificate Examination in 1997. The 12-point scale associated with the Junior Certificate Examination reported in studies prior to PISA was actually an average across subjects. The 12-point performance scale reported in Shiel et al. and Cosgrove et al. is for individual subjects (English, mathematics, and science). The scale has a three-grade overlap between syllabus levels, such that an A at higher level corresponds to 12 points, at ordinary, 9, and at foundation, 6.

Martin and Hickey (1992) noted that the points schemes for Leaving Certificate used for selection for third-level education suffer from a common problem in that they were designed to discriminate between candidates with high levels of performance, and give little credit (and hence, allow comparatively less discrimination) for grades at ordinary level (this was also discussed in Chapter 1). One could argue that the problem of the points scheme that applies at Leaving Certificate also applies at Junior Certificate, as a second, earlier gateway to selection to third level, since the majority of students taking a subject at a syllabus level either retain the same syllabus level at Leaving Certificate, or switch to a 'lower' level (see, for example, Millar & Kelly, 1999, pp. 56-57).

Associated with Martin and Hickey's general observation, Millar and Kelly (1999) identified two specific problems with the conversion of letter grades to a numeric scale. First, the scores are affected by the syllabus level at which the examination is taken. Second, the scale assumes that the distances between scale points are equal, even at the extremes. Millar and Kelly speculate that the scale may be stretched at the lower end of higher and ordinary levels, and more closely clustered at Foundation level. Similarly, Kellaghan and Dwan (1995) comment that the weights assigned to grades are "somewhat arbitrary" (p. 17).

A further problem with the scale, which is particularly relevant to analyses involving the PISA cohort, is that there are very few students in the foundation-level group (in English in particular), and most of these are clustered at the upper end of the letter grades.

Millar and Kelly (1999) comment: "It would be preferable to have an independent measure, if such were available, against which to compare the performance of candidates across all subjects and levels" (p. 227). The PISA achievement data provide

an independent measure of sorts, but establishing linkages between the two assessments is complicated by the fact that some students participating in PISA took the Junior Certificate in the year prior to PISA. The analyses presented in Shiel et al. (2001, p. 224) reported the results of different ways of scaling the JCPS before selecting the 12-point scale for analysis of the PISA 2000 data (the same 12-point scale was used once again in Cosgrove et al., 2005). Different amounts of overlap across the syllabus levels were examined by computing the Pearson correlation coefficient between Junior Certificate English/PISA reading scale scores and Junior Certificate mathematics/PISA mathematics scale scores. In the case of Junior Certificate English, 10-point and 14-point scales were compared to the original 12-point scale. In the case of mathematics, 10-, 14-, and 16-point scales were compared. For English/reading, correlations were .729, .742, and .737 for the 10-, 12-, and 14-point scales, respectively. Correlations for mathematics were .703, .729, .730, and .725, for the 10-, 12-, 14-, and 16-point scales, respectively. These analyses, however, did not examine different ways of scaling the Junior Certificate letter grades to take the stretching/clustering noted by Millar and Kelly (1999) into account; nor did they consider ways of treating the very low numbers of students who took the examination at Foundation level and received grade C or lower.

Chapter 4, therefore, probes the interpretation of Irish student performance on both PISA reading and mathematics further by comparing various ways of scaling Junior Certificate data so as to produce a best match with achievements on the PISA tests, both for the overall scales, and on the subscales. Findings are then related back to the existing research.

A second aspect of addressing the question as to what PISA can tell us about the achievements of students in Ireland is a consideration of the extent to which PISA is capable of describing not only average achievement, but also the achievements of students at the extremes of the achievement distribution, and of subgroups of policy interest. Thus, the dissertation also entails an analysis of bias in the achieved PISA samples and considers the potential consequences and limitations imposed on the results, were it to be shown that the PISA samples for Ireland were significantly biased. Bias arising from non-response in particular (as opposed to other sources of bias, such as low coverage of the population) has been identified as problematic. This theme is

explored in depth in Chapter 3, where school and student non-response is analysed using logistic regression to investigate whether or not we can be confident about conclusions about Ireland's performance, and the extent to which estimates of the performance of particular subgroups of the population may be considered reliable.

2.6.2. What Does PISA Tell us About the Equity of Achievement Outcomes in Ireland?

The literature review has noted the high emphasis placed by the OECD and in the media on 'educational equity', which is billed as a desirable outcome of education systems. Usually, the percentage of achievement variance which is between schools is used as an indicator of educational equity, particularly by the OECD.

However, it has been shown that conclusions about the relative equity of an education system may vary depending on whether the sampled design entails a random within-school sample on the basis of age, or the sampling of intact classes, since these give rise to differences in the amount of achievement variance that is between 'schools', particularly if class allocation is based on ability. There is evidence that these differences apply in the Irish context from a review of the research on achievement outcomes in Ireland. Chapter 5 examines this issue further by comparing the variance components associated with PISA and TIMSS; it also includes a re-analysis of the TIMSS data which partitions achievement into three components (school, class, student).

Between-school variance may also vary as a function of the extent to which an achievement measure is intended to assess the curriculum, with higher between-school variance being associated with more curriculum-sensitive measures, although recent data are lacking. This suggests, in the case of PISA mathematics in particular, that between-school variance should not be interpreted to reflect between-school differences in variables relating to the content and delivery of the mathematics syllabus. The literature review of achievement variance in Ireland also suggested that magnitude of between-school variance may rest in part on the school-dependence of the subject. More generic skills such as reading, whose components are learned and practised in many contexts outside of formal schooling, may be less sensitive to between-school differences in instructional etc. variables. These aspects of the achievement measure are

also explored in Chapter 5, where the variance components of PISA reading, Junior Certificate English, PISA mathematics and Junior Certificate mathematics are compared.

It is unfortunate that PISA did not include a grade-based sample (of intact classes) alongside its age-based sample, since this would have allowed a much more robust investigation of the issues. Nonetheless, the results in Chapter 5, which capitalise on the survey design differences between PISA and TIMSS, may be considered a first step in disentangling the extent to which both achievement measure and sample design should be borne in mind when drawing inferences about the equity of systems.

Chapter 3 also considers whether school and student non-response has an impact on total and between-school variance in Ireland in order to ascertain whether conclusions drawn about the relative homogeneity of schools and students are reliable on the basis of non-response.

2.6.3. What Does PISA Tell us About the Determinants of Achievement in Ireland?

Considerable emphasis is placed on explanatory models of achievement in both the international and national reports on PISA and other surveys of educational achievements, and these have tended to focus on the extent to which social background and school/class practices explain achievements.

It has been shown, though, that the curriculum sensitivity of the test and the extent to which it may be considered school-dependent may be related to the extent to which social background and school/class practice variables may impact on achievement. This is important since conclusions one might draw about the relative importance of SES and school/class practices may vary, depending on the nature of the test measure.

However, it is not sufficient to consider the content of the test in isolation; it is also necessary to consider how the sample was selected. It has been shown in the material reviewed in Chapters 1 and 2 that age-based and grade-based samples are both associated with advantages and disadvantages, and have features which should be taken into account in the interpretation of achievement results, particularly how they apply to

achievement variance. The main advantage of a grade-based sample is that it allows one to pin outcomes to a particular point in the system. However, particularly if only one class is sampled per school, the sample may not be considered to be representative at the school level, if students are allocated to classes on a non-random basis. Results arising from an age-based sample, in contrast, make it difficult for results to be pinned to a particular point in the system if students are dispersed across multiple grade levels, but, on the other hand, the sample is representative of the achievements of individual schools to a greater degree than a grade-based sample such as used in TIMSS. Sample design has important potential consequences for the interpretation of the relative impact of social intake on achievement and school/class-level variables on achievement, whereby grade-based samples might inflate the impact of these, and age-based samples underestimate them, although these phenomena have not been the subject of research in Ireland for close to 30 years. Thus, a consideration of the impact of the sample design in conjunction with the nature of the test measure in explanatory models of achievement forms the third broad theme of this thesis. The theme is explored in Chapter 5 using multilevel models which compare achievement on PISA 2000 reading, 2000 Junior Certificate English, PISA 2003 mathematics, 2003 Junior Certificate mathematics, TIMSS 1995 mathematics, and 1996 Junior Certificate mathematics. Again, it is unfortunate that PISA did not include a grade-based sample alongside an age-based sample to allow for stronger inferences to be drawn about these issues, so the analyses in Chapter 5 should be regarded as an initial step rather than conclusive or definitive.

Table 2.14 *Outline of the Three Features of PISA Examined in the Thesis and Their Consequences for Interpreting Results*

Is the sample biased or not?		*Is the achievement measure similar to, or different from, what is assessed nationally?*		*Is the sample design grade- or age-based?*	
Biased	*Not biased*	*Similar*	*Different*	*Grade*	*Age*
Depending on the nature of the bias, one cannot be confident about the achievement levels of some subgroups of policy or theoretical interest; If lower response rates are associated with lower achievers then claims about the extent to which literacy problems prevail are not accurate (and the monitoring of trends/ setting of targets is, as a corollary, also problematic); Achievement variance may be underestimated, resulting in incorrect conclusions about the homo/heterogeneity of the sample surveyed and possibly erroneous conclusions about the relative 'equity' of the system	*Provided other sources of bias are controlled for:* One can be confident about the achievement levels of subgroups of policy or theoretical interest; Claims about the extent to which literacy problems prevail are accurate and the data may be used with confidence as a basis for monitoring literacy problems/setting targets; Achievement variance is unbiased, resulting in correct conclusions about the homo/heterogeneity of the sample surveyed and allowing claims about the 'equity' of a system to be made on a firm basis	National achievement measures are in line with international measures; The assessment can be used in a straightforward manner to draw conclusions about the education system (utility and interpretability of the survey results are good)	National achievement measures are at odds with international measures; The interpretation of outcomes is not straightforward (a consideration of how national and international measures differ is needed); It is difficult to draw conclusions about the education system (utility and interpretability are limited)	Allows one to 'pin' outcomes to a specific point in the education system (which is potentially useful for policy development); Sample may be considered representative of the population but not of individual schools (unless all classes at a grade level are selected); Estimates of between-'school' achievement variance are inflated since they are confounded with between-class variance, which is likely to be based on ability in Ireland; Interpretation of the nature and size of the 'social context effect' of the school is confounded with class allocation	Does not allow one to 'pin' outcomes to a specific point in the education system if sampled students are dispersed across grade levels/programmes of study; Sample may be considered representative both of the population and of individual schools; Partitioning of achievement variance is not confounded with class allocation, but may disguise large achievement differences within schools; Interpretation of the nature and size of the 'social context effect' of the school is not confounded with class allocation
Additional Considerations:		*Additional Considerations:*		*Additional Considerations:*	
The efficiency of non-response adjustment varies according to the amount of variance between 'schools'		More school-dependent subject areas may be more sensitive to 'school effects' (variables relating to school policies and practices) than school-independent ones; Curriculum-sensitive measures may be more sensitive to school effects than curriculum-neutral ones		The manner in which social background is measured is likely to be suitable for policy development or theoretical development, but not necessarily both	

2.7. Conclusion

This chapter considered PISA in the Irish context with respect to four aspects. First, the potential for bias in estimates of mean achievement and variance in achievement arising from non-response with respect to the Irish PISA datasets was considered. Second, existing research on PISA and the curriculum in Ireland was reviewed with the aim of ascertaining what the PISA measures can tell us about achievement in Ireland, to identify further analyses, and to consider whether the design of PISA itself might be improved in this regard. Third, a review of published analyses of between-school/class variance in the achievements of Irish students was undertaken to identify patterns in the results which may be linked to the sample design and/or the nature of the achievement measure. Fourth, explanatory models of achievement in Ireland were also reviewed in order to develop a number of hypotheses as to how test content and sample design should be considered when interpreting results of explanatory analyses.

Regarding the first issue, there is some evidence from existing research that student non-response in particular (as opposed to non-response at the school level) may be associated with an upward bias in mean achievement, given that the between-school variance in achievement in Ireland is comparatively low. Further, there is a possibility that variance in achievement (both total variance and between-school variance) are downwardly biased as a result of non-response (since, if the non-responders are at the extremes of the achievement distribution, the obtained distribution will be more homogenous than would have been the case had all eligible students participated). However, this research is based on simulated datasets and does not address the question as to whether some policy-relevant subgroups in Ireland might also be under-represented.

The literature review regarding the second theme focused primarily on existing research which compared the PISA tests and national curricula (specifically, the Junior Certificate Examinations) (Close & Oldham 2005; Cosgrove et al., 2005; Shiel et al., 2001). The comparison of Junior Certificate English/PISA 2000 reading indicates that, while the tests differ in format, relative emphasis on literary/functional texts and the manner in which students' responses are marked, the reading skills assessed are quite similar, particularly for higher- and ordinary-level students. Thus one cannot argue that

PISA reading is a curriculum-free measure; rather that PISA provides an assessment which is compatible with the national curriculum (if providing an incomplete assessment of it).

Considerable differences between the style of the PISA 2003 mathematics test and Junior Certificate mathematics were found which can be attributed to differences in the underlying philosophies of the assessments. PISA mathematics is based on the Realistic Mathematics Education movement, which emphasises horizontal mathematisation in concrete and authentic contexts. Junior Certificate mathematics takes a more formal and abstract approach, with a considerable number of theorems, proofs, and vertical mathematisation processes.

Some limitations of the analyses were noted, perhaps the most significant of which is that they do not explain student achievement once achievement on the Junior Certificate is taken into account. Although not an optimal method of adjusting the relationship between curricular familiarity and achievement this finding is consistent with previous research. Measures of curriculum familiarity in cross-sectional designs may be inherently confounded with student ability (Floden, 2002). This suggests that the test-curriculum rating project results are better suited to providing a broad profile of how PISA and the Junior Certificate overlap and diverge rather than as an explanation of the achievements observed. A second major limitation is that these results cannot be compared to those of other countries, so it cannot be known whether an aspect of the PISA test which rated as having low familiarity to students in Ireland might in fact be relatively familiar, given the curricula of other countries. Third, other aspects of the assessments not considered in the analyses may also be relevant. Finally, the analyses give no information on how the curriculum has been implemented. They yield broad information on the intended curriculum only.

A review of the performance of Irish students on PISA 2000 reading and PISA 2003 mathematics indicates that in international terms, reading standards are comparatively high, and that there are fewer low achievers in Ireland. In mathematics, performance is only around the OECD average, and there is variability in standards when one considers performance on the four mathematics subscales. In both reading and mathematics, the dispersion of achievement was comparatively narrow and between-school variance

comparatively low. In OECD terms, therefore, Ireland is a 'high-equity' country, and also high-achieving in the case of reading.

Substantial achievement differences were observed across the three syllabus levels in both reading and mathematics. From a comparison of the results for reading and mathematics, one might argue that the 'average' student at a syllabus level in mathematics has a slightly higher standard of achievement than the 'average' student in English in national terms. However, when one compares the mean scores of higher-, ordinary- and foundation-level students to the OECD average for reading and mathematics, it is apparent that, in international terms, higher-level students are 1.6 standard deviations above the OECD average in both subjects.

The distribution of higher-, ordinary- and foundation-level students across PISA 2000 reading proficiency levels and PISA 2003 mathematics proficiency levels indicates some commonalities across the two subject areas. First, a substantial proportion of foundation-level students who attempted PISA did not demonstrate sufficient skills to be placed reliably on the scale in question; a second substantial proportion were placed at the lowest proficiency level, able to demonstrate only the most basic literacy skills assessed in PISA. Further, a substantial proportion of ordinary-level students are at or below Level 1 in both reading and mathematics. However, the differences in the percentages of students taking foundation-level English and foundation-level mathematics who score at or below Level 1 on PISA reading and mathematics, respectively, suggests that taking foundation-level English is associated with particularly weak achievement. At the upper end of the proficiency distributions in both subject areas, a substantial minority of students taking ordinary level (about one in 10 in the case of English reading and one in eight in the case of mathematics) are achieving proficiency levels 4 or higher.

Disparities between achievement on PISA and the Junior Certificate are evident when one compares the percentages of ordinary- and foundation-level students in particular who are at or below Level 1 in reading and mathematics with the percentages of students awarded below a grade D on the corresponding Junior Certificate subject. Clearly, the Junior Certificate and PISA serve different purposes and 'low achievement' on one assessment is not the same as 'low achievement' on the other; however, that large

proportions of ordinary- and foundation-level students have only the most basic (or even below basic) literacy skills is made clear from these analyses. If the position of the OECD (2001b; 2004c) with respect to being at or below Level 1 is correct (i.e., that individuals at these levels have inadequate skills to build on for successful participation in adult society), then this is a cause for concern. While the Chief Examiners' reports suggest that some students appear to struggle with the examination and are not of the expected standard, the PISA results suggest, in particular, that a substantial minority of *ordinary*-level students are below the minimal standards required for future educational, personal and occupational requirements.

A comparison of the percentages achieving various grades in both English and mathematics with the percentages at each proficiency level in the case of higher-level students suggests that the percentages at each proficiency level are more closely aligned to the percentages of higher-level students attaining each letter grade than those at ordinary or foundation level, and that the letter grades associated with higher-level students may be amenable to interpretation according to the standards implied by the PISA proficiency levels.

The large disparities in achievement, together with the apparent mismatch between ability (as measured by PISA) and Junior Certificate syllabus level taken, calls into question the appropriateness of the syllabus level taken by some students. Unfortunately, there is no published research which examines the processes whereby students come to select (or are selected to take) a particular syllabus level; syllabus documentation is also scant on concrete guidelines for methods of identifying the optimal syllabus level for students of differing abilities and needs. Knowing how and why students come to take the syllabus levels they do may shed some valuable insights into the PISA-Junior Certificate achievement mismatch observed in the data.

Although performance on PISA and the respective Junior Certificate subject indicates a moderate degree of overlap, correlations between PISA mathematics and PISA reading, and between Junior Certificate mathematics and Junior Certificate English, are of a similar magnitude (around .70 in all cases). This suggests that characteristics of the tests (e.g., item formats), and the circumstances under which the tests are taken (e.g., high-stakes in the case of the Junior Certificate versus low-stakes in the case of PISA) are

relevant in considering these results. Another issue is that while existing research has explored links between performance on PISA and the Junior Certificate, it is not clear whether the 12-point JCPS used in published comparisons should be modified to account for possible clustering at the lower end of the achievement scale and stretching at the upper ends. Furthermore, were differences to be found between Junior Certificate English and Junior Certificate mathematics in these respects, they may help further with an understanding of what PISA tells us about student achievement in Ireland.

Regarding the issue of curriculum, it might be noted that the results for PISA mathematics have given rise to more discussion and commentary (e.g., Close and Oldham, 2005; Close et al., 2005; Oldham, 2002) than those of PISA reading, and the mathematics curriculum was the focus of much of the discussion at the PISA 2003 national symposium (Educational Research Centre, 2005). Further, in Chapter 1, it was noted that media reports of the PISA results tended to be uncritical in the case of the English curriculum, but that there were calls for a review of the mathematics curriculum. Differences in how the PISA results in reading and mathematics have been received have arisen as a result of disparities in the relative performance of Irish students on PISA reading and PISA mathematics in both 2000 and 2003; they have also arisen on the basis of wide divergences in the philosophies underlying PISA mathematics and Junior Certificate mathematics, the topic areas assessed, and even more so, the style of the assessments. In this sense one could argue that the PISA survey has had a stronger impact on mathematics education in Ireland than on English/reading education. In fact, a substantial review of mathematics education at post-primary level has been undertaken by the NCCA (2005). In the review, PISA is mentioned on two occasions – once to illustrate that mathematics education in Ireland is at odds with the Realistic Mathematics Education movement underpinning PISA and evident in curricula elsewhere, and also to illustrate, in a more general discussion on concerns about mathematics standards, that Irish performance in international terms is only around the OECD average on PISA, in contrast with TIMSS, where it was above average. A suggested reason for this is the mismatch between the PISA mathematics test and the Irish mathematics curriculum at Junior Cycle.

To address the third issue, between-cluster variance in achievement in a number of studies of achievement in Ireland was reviewed. It was noted that, when the class is

taken as the unit of analysis, between-cluster variance tends to be higher, particularly for standardised (or curriculum-insensitive) measures which suggests that factors relating to class allocation are related to student ability. It was also noted that between-school variance tends to be higher for curriculum-sensitive measures, particularly mathematics and Irish, which is indicative of differential school effectiveness in the implementation of these curricula. Between-school variance appeared to be lower for less curriculum-dependent subjects such as English (although it was noted that the research on which these inferences are based is now over 25 years old). A comparison of variance components for TIMSS 1995 and PISA 2000 suggests that between-class variance is comparatively high in Ireland, while between-school variance is comparatively low; however, differences in the test content should be taken into account here. A comparison of the variance components for PISA 2000 and students taking Junior Certificate English, mathematics and science in 1999 or 2000 suggests that Junior Certificate mathematics is associated with slightly higher between-school variance than PISA mathematics; however, that only slight differences were evident may be partly a function of PISA's age-based sample design (a more marked difference might have been found, had PISA sampled intact classes). Unfortunately, the design of PISA does not allow one to simultaneously assess the impact of the test measure and the sample design on variance components and achievement (this would only be possible had PISA employed a hybrid age- and grade-based design, which was the case in FIMS, reviewed in Chapter 1).

The final area considered the results of several studies which reported explanatory models of achievement in Ireland. A study by Madaus et al. (1979) which examined the unique contributions of class-level and other variables to curriculum-sensitive and standardised measures of achievement is key to the development of the research questions which are explored in Chapter 5. However, the analysis methods suffered from some limitations, and the results may not apply to the same degree in the present, given changes in the education system (e.g., increased enrolment; curricular revisions). This is taken as an indication (together with the high emphasis in recent international and national reports on PISA placed on explanatory models particularly in terms of discussions of equity and the impact of school/class variables), that it is time to revisit the issues.

It was pointed out that comparisons across explanatory analyses of achievement are difficult due to differences in the achievement measures, sample design, and explanatory variables used. However, there is consistent evidence for a moderate to strong impact of social intake in Irish schools, whereby the social mix of the school exerts an influence on achievement over and above the individual student's socioeconomic status or social background. Further, this effect appears to be linear and there is also some evidence of a cross-level interaction between social context and student gender, whereby the effect of the social context is stronger for boys. This effect does not appear to be limited to a single subject area, since it has been detected for both English and mathematics, as well as for composite measures of performance. Models which do not include an adjustment for prior ability (intake) of students tend to explain the majority of achievement variance between schools, but a third or less of the variance within schools. When intake measures are included, the percentage of within-school variance increases substantially; however, there is no study of achievement in Ireland which has included an intake adjustment which was actually measured at the time of intake.

More recent multilevel models which compared achievement on PISA 2000 and on 1999/2000 Junior Certificate English represent more recent attempts within a multilevel modelling framework to compare curriculum-sensitive and non-sensitive measures. In contrast to Madaus et al. (1979), the models are quite similar to one another. Gender interactions with attitudinal variables and home background were slightly more prevalent in the Junior Certificate models. Since these models did not partition out the variance attributable to student and school social background from other variables, however, it is impossible to say whether the models differ in this regard or not. Further, since students are dispersed across three grade levels, and took the Junior Certificate in two different years, direct comparisons between the models are complicated. A third problem is the lack of class-level and school-level variables included in the models. A final problem is that the analyses do not allow for comparisons of the impact of sample design on the results. Analyses presented in Chapter 5, therefore, are better designed to address the issues raised by Madaus et al.

CHAPTER 3. ANALYSES OF NON-RESPONSE BIAS IN IRELAND ON PISA

3.1. Introduction

The analyses reported in this chapter consider how the obtained sample of students and the statistical adjustments for non-response may introduce bias in the results for Ireland. They are designed to answer the question as to what PISA can tell us about the achievement of students in Ireland, particularly achievement at the extremes of the distribution, and of particular subgroups of the population.

Results of analyses are reported. These attempt to determine

(i) whether schools that did not respond to the PISA 2000 and PISA 2003 assessments differ in certain respects from schools that responded,

(ii) whether students that did not participate to the PISA 2000 and PISA 2003 assessments differ in certain respects from students that participated,

(iii) whether the non-response of schools results in a downward bias in total variance in achievement,

(iv) whether the non-response of students results in a downward bias in between-school variance in achievement.

3.2. Rationale

The analytic procedure for (i) and (ii) takes the form of a series of logistic regressions at the level of the school and at the level of the student. The outcome variable for both the school and student analyses is participation status (participated/did not participate) and the explanatory variables pertain to various aspects of the schools (e.g., designated disadvantaged status, school type) and of the students (e.g., student gender, syllabus level at which the Junior Certificate was taken). The procedure is considered appropriate since a similar approach was taken by the PISA consortium in cases where school-level response rates fell below 85% (Adams, Rust, & Monseur, 2002).

Logistic regression was chosen both because of the binary nature of the outcome variable and because it allows the evaluation of the effects of multiple explanatory variables (and their potential interactions) simultaneously (Hutcheson & Sofroniou, 1999, pp. 113-161). However, in contrast to the approach taken by the PISA consortium who analysed school-level non-response only, analyses are carried out at both school

and student levels, since the sample design is in two stages, entailing selection of schools, then students.

Characteristics of responding and non-responding schools are compared, followed by logistic regression analyses of school response status. The mean scores of five groups of responding and non-responding students on the Junior Certificate examination (English in the case of PISA 2000 and mathematics in the case of PISA 2003) are then compared (present, absent, refused, left school, special educational need) in a one-way ANOVA. Characteristics of eligible responding and non-responding students (i.e., those who have not left the school and have not been identified as having a special educational need) are compared next, followed by logistic regression of student response status.

The analytic procedure for (iii) and (iv) compares, within a multilevel modelling framework, the between-school and total variance components for the achievements of students who did participate in PISA 2000 and PISA 2003, with all available achievement data for students, both participating and non-participating, for PISA 2000 and PISA 2003. Since PISA achievement data is not available for non-participating students, the comparisons are made on the basis of Junior Certificate examination results. The analyses assume, therefore, that any differences in the variance components arising from comparisons of Junior Certificate data would also have occurred with the PISA measures of achievement.

3.3. Comparison of Responding and Non-Responding Schools in the PISA 2000 and PISA 2003 Samples for Ireland

Table 3.1 provides a comparison of school characteristics by school participation status for 2000 and 2003. Participating schools include a small number of 'replacement' schools, which are identified automatically by the sampling procedure (as described in Chapter 1).

Table 3.1. Cross-Classification of School Participation Status with School Type, Sex Composition, and Designated Disadvantaged Status: PISA 2000 and PISA 2003

	PISA 2000				PISA 2003			
	Did not participate		Did participate		Did not participate		Did participate	
Characteristic	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*
School Type								
Secondary	9	60.0	87	62.6	9	100	83	57.2
Vocational	2	13.3	33	23.7	0	0	35	24.1
Community/Comprehensive	4	26.7	19	13.7	0	0	27	18.6
School Sex Composition								
All boys	5	33.3	22	15.8	3	33.3	28	19.3
All girls	3	20.0	35	25.2	1	11.1	33	22.8
Mixed sex	7	46.7	82	59.0	5	55.6	84	57.9
School Designated Status								
Designated disadvantaged	11	73.3	104	74.8	1	11.1	41	28.3
Not designated disadvantaged	4	26.7	35	25.2	8	88.9	104	71.7
Total	15	100	139	100	9	100	145	100

Since the replacement schools are comparable to the original schools in terms of explicit (school enrolment size) and implicit (school sector and school sex composition) stratification variables, the present analysis treats participating schools, whether original or replacement, as equivalent. In 2000, there is slight under-representation of community/comprehensive schools and of all boys' schools, but otherwise the participating and non-participating groups are similar. In 2003, all nine non-participating schools are in the secondary sector. The participation rate of all boys' schools was somewhat higher than all girls' or mixed sex schools, while participation rates for designated disadvantaged schools was a little higher than in those not designated.

3.4. Logistic Regression of School Participation Status

A series of logistic regressions with response status of the school (participate/did not participate) as the outcome were carried out to examine whether the likelihood of participation differed according to school characteristics in PISA 2000 and PISA 2003. Dummy coding of categorical variables was implemented where appropriate. This results in a series of binary indicators, with one of the categories of the variable designated as the 'reference' category, to which parameters of the other categories are compared (see Hutcheson & Sofroniou, 1999, pp. 85-90).

3.4.1. Variables

The following were used as explanatory variables in both 2000 and 2003:

- school size, as indicated by the explicit sample stratum (categories: small, medium and large. Since all selected small schools participated in both 2000 and 2003, these were collapsed into a single category with medium schools and the comparison is between large schools (the reference category) and other schools)
- school designated disadvantaged status (categories: yes, no; reference category = yes)
- school sex composition (categories: all boys, all girls, mixed sex; reference category = mixed sex)[29]
- school type (categories: secondary, community/comprehensive, vocational; reference category = secondary).
- School mean performance as indicated by the mean Overall Performance Score (OPS) on the 1999 Junior Certificate of *all* Junior Certificate candidates in those schools 1999 (continuous, the summed score for students' highest seven subjects in the Junior Certificate aggregated to the level of the school; PISA 2000 range = 41.1 – 75.3, M = 64.7, sd = 6.15; PISA 2003 range = 41.4 – 75.3, M = 64.8, sd = 6.26). These data were extracted from the 1999 Junior Certificate examinations database, aggregated to the school level, and matched to the PISA 2000 dataset using school roll number. In the case of PISA 2003, more recent data were not available (a request for more recent data sent to the Post-Primary Administration Section of the Department of Education and Science on April 25, 2005 was not successful), so the analysis rests on the assumption that any changes to Junior Certificate performance from year to year are not biased in a particular direction with respect to the schools selected to participate in PISA. This would appear to be reasonable, given that PISA schools are sampled at random (using probability proportional to size).

[29] In the analyses of non-response at the student level described later in this chapter, school sex composition is coded as a binary variable (mixed sex/single sex) since the student gender is included as a variable in the student-level analyses of non-response. Also, there are slight differences in the school sex composition as used in PISA 2000 and PISA 2003. In PISA 2000, the sex composition was a classification taken from the Department of Education and Science's schools database; in 2003 it was computed based on the actual number of female 15-year-olds enrolled in each school according to the sampling frame. For ease of comparison, however, the Department of Education and Science's classification is used in both sets of analyses.

3.4.2. Procedure

Following model-building procedures similar to those used in the multilevel models described in the PISA 2000 and 2003 national report for Ireland (Shiel et al., pp. 98-99; Cosgrove et al., 2005, pp. 137-139), each explanatory variable was first tested separately and its significance evaluated using the Wald chi-square statistic. The Wald statistic indicates the improvement in model fit after addition of each variable (Hutcheson & Sofroniou, 1999, p. 139). The analyses are unweighted.

3.4.3. Results: PISA 2000

Table 3.2 shows the outcome of a series of logistic regressions testing each variable separately for the PISA 2000 dataset.

Table 3.2. School-Level Binary Logistic Regressions with PISA 2000 School Participation as the Outcome Variable, and Five Explanatory Variables Tested Separately

Variable/Comparison	*B*	*SE (B)*	*df*	*p*	*Exp (B)*
School Size (Sample Stratum)					
Large stratum – Small/Medium stratum	-0.140	0.613	1	.819	0.869
Designated Status of School					
Designated - Not Designated	-0.077	0.616	1	.900	0.925
School Sex Composition					
Mixed Sex – All Boys	-0.979	0.633	1	.122	0.376
Mixed Sex – All Girls	-0.004	0.719	1	.995	0.996
School Type					
Community/Comprehensive - Secondary	-0.711	0.652	1	.276	0.491
Vocational - Secondary	0.535	0.808	1	.508	1.707
School Mean Achievement	-0.038	0.050	1	.446	0.963

Note. Data are unweighted.

For the four categorical variables, there is no difference in the likelihood of participation across groups, and the mean achievement of schools that did and did not participate is the same ($t = 0.762$; $df = 152$; $p = .447$); in fact, the mean achievement on the 1999 Junior Certificate of non-participating schools ($M = 65.9$; $SD = 5.78$) is marginally higher, by about $^1/_5$ of a standard deviation, than that of participating schools ($M = 64.6$; $SD = 6.20$).

Since none of the variables is significant, it was not necessary to proceed with constructing a model examining these variables simultaneously. It can be concluded that the PISA 2000 school sample may be regarded as representative in terms of these particular variables.

3.4.4. Results: PISA 2003

Table 3.3 shows the results of a series of logistic regressions testing each variable separately for the PISA 2003 dataset. School type (whether secondary, vocational, or community/comprehensive) was not included as a variable in Table 3.3 since there is zero variance in the response variable for two of the three school types. Even if two of the three school types were collapsed into a single group, this would still result in a group with zero variance on the response variable. Logistic regression models cannot compute coefficients for variables with zero variance on one or more variables. However, a chi-square test comparing secondary schools with 'other' school types is significant (χ^2 = 6.442, df = 2, p = .011), indicating a significant association between type and participation status, whereby secondary schools are less likely than the other school types to participate. There is no difference in the participation likelihood of schools in PISA 2003 by any of the other variables examined. Consistent with PISA 2000, the mean achievement on the 1999 Junior Certificate of non-participating schools (M = 67.2; SD = 4.45) is a little higher, by about $^2/_5$ of a standard deviation, than that of participating schools (M = 64.6; SD = 6.34).

Table 3.3. School-Level Binary Logistic Regressions with PISA 2003 School Participation as the Outcome Variable, and Four Explanatory Variables Tested Separately

Variable/Comparison	*B*	*SE (B)*	*df*	*p*	*Exp (B)*
School Size (Sample Statum)					
Large stratum – Small/Medium stratum	0.042	0.729	1	.954	1.043
Designated Status of School					
Designated - Not Designated	1.149	1.077	1	.286	3.154
School Sex Composition					
Mixed Sex – All Boys	0.558	1.116	1	.617	1.747
Mixed Sex – All Girls	-0.412	0.759	1	.588	0.663
School Mean Achievement	-0.084	0.071	1	.235	0.919

Note. Data are unweighted.

Therefore, with the exception of the lower participation rate of secondary school types, the PISA 2003 school sample may be regarded as representative. Since the school non-response adjustment includes adjustments for implicit strata (which includes school type), the PISA 2003 school sample may be regarded as being free from non-response bias on the variables examined in the model.

3.5. Comparison of Responding and Non-Responding Students in the PISA 2000 and PISA 2003 Samples in Ireland

3.5.1. Types of Non-Response

For the purposes of this analysis, five groups of students are distinguished:

1. Participating students
2. Absent students, no known reason
3. Refusing students
4. Students with a known special education need
5. Students no longer in the school.

The age-ineligible students (18 in 2000 and 4 in 2003) have been excluded from the analysis since they should not have been included on the original list of sampled students. The numbers of students in each of these categories for PISA 2000 and PISA 2003 are shown in Table 3.4.

Table 3.4. Numbers (and Percentages) of Students in PISA 2000 and 2003, by Response Status

	PISA 2000		*PISA 2003*	
Student Participation Group	*N*	*%*	*N*	*%*
Participated	3854	80.4	3880	77.7
Absent	582	12.1	761	15.2
Refused	119	2.5	91	1.8
Left School	102	2.1	139	2.8
Special Educational Need	134	2.8	125	2.5
Total	4791	100	4996	100

Note. The table does not include age-ineligible students (18 in PISA 2000 and 4 in 2003) who were included in the databases in error.

Overall, the figures are similar for the two years, although the rate of absenteeism is slightly higher in 2003 than in 2000. As noted in Chapter 2, the weighted student-level participation rates are above the required minimum standard of 80% in both 2000 and 2003.

3.5.2. Matching the PISA Student Datasets with the Junior Certificate Examinations Datasets

To conduct analyses reported in the PISA 2000 national report for Ireland (Shiel et al., 2001, Chapter 6), the student dataset was matched with the Department of Education and Science Junior Certificate Examinations databases for 1999 and 2000. About one-third or 32.6% (n = 1,257) of PISA 2000 students took the examination in 1999 and

61.2% (n = 2,360) of students took the examination in 2000[30]. The match was obtained by creating a combined school / sex / grade level / date of birth identification code and running an automated match. Where no match was observed in this process, it was obtained by manually comparing student name on both databases. A match for 94.1% (n = 3,625) of students was obtained.

However, for the purposes of the present analysis, it was necessary to include Junior Certificate results (using the same method) for *all* sampled PISA 2000 students, both those that did participate and those that did not. Once this was carried out, using the same procedure, of the 4,791 students selected, a match was obtained for 4,344 (90.7%). Since the match entailed Junior Certificate Examinations data from 1999 and 2000 only, Junior Certificate information is not available for students who were in Second year, or in Fifth year (following completion of the Transition Year Programme) at the time of the PISA 2000 assessment.

The match between the PISA 2003 student dataset and the Junior Certificate results for 2002 and 2003 was made using a similar procedure; the only difference was that the match was made in one step: all sampled PISA 2003 students were subjected to the matching procedure rather than just those students that actually participated in PISA 2003. Of participating students, about one-third or 34.4% (n = 1,333) took the examination in 2002 and 59.7% (n = 2,315) of students took the examination in 2003. An overall match for 93.9% (n = 3,645) of the participating students was obtained. Of all sampled students, a match was obtained for 4,553 (90.1%).

The manner in which performance across different syllabus levels on the Junior Certificate examination was placed on a single 12-point Junior Certificate Performance Scale (JCPS) is shown in Table 3.5 and described in more detail in Chapter 4.

[30] These figures differ very slightly from those reported in Shiel et al. (2001, pp. 145-146) because unweighted figures were reported in error in that report. The figures reported here are weighted.

Table 3.5. Mapping of Junior Certificate Letter Grades and Syllabus Levels onto Junior Certificate Performance Scale (JCPS)

Syllabus Level	*Junior Certificate Performance Scale Score*											
	12	11	10	9	8	7	6	5	4	3	2	1
Higher	A	B	C	D	E	F						
Ordinary				A	B	C	D	E	F			
Foundation							A	B	C	D	E	F

Note. Students obtaining 'No Grade' receive no credit on this scale. Across all syllabus levels, letter grades are awarded based on percent correct: A = 85%+, B = 70-84%, C = 55-69%, D = 40-54%, E = 25-39%, F = 10-24%, NG = <10%.

3.5.3. Performance of Students on the Junior Certificate by Participation Status

Table 3.6 shows the mean Junior Certificate English performance scores (EJCPS) for PISA 2000, and Table 3.7 shows the mean Junior Certificate mathematics performance scores (MJCPS) for PISA 2003, for the five groups of interest (present, absent, refusal, special educational needs and left school).

Table 3.6. Mean Junior Certificate English Performance Scores (EJCPS) for the PISA 2000 Cohort, by Participation Status

Group	*N total*	*N available*	*% available*	*EJCPS Mean*	*SD*	*SE*
Present	3854	3625	94.1	9.19	1.71	0.028
Absent	582	489	84.0	8.41	1.90	0.086
Refusal	119	106	89.1	8.28	1.98	0.193
Special Needs	134	79	59.0	6.72	2.25	0.253
Left School	102	45	45.1	6.67	1.95	0.291
Total	4791	4344	90.7	9.01	1.81	0.028

Note. Estimates are unweighted. The standard errors reported in this table are somewhat underestimated since the clustered nature of the sample design has not been taken into account in the analysis. The total N does not include 18 age-ineligible students.

Comparing the percent of available cases across 2000 and 2003, the overall pattern is similar, with the lowest rate of available data for the special needs and left school groups. The percentages for these two groups are a little higher in 2003, however. Caution is advised in interpreting the mean scores of the special needs and left-school groups in Tables 3.6 and 3.7 since the match rate is quite low (less than 65%). It is likely that the majority of the unmatched students in these two groups either (i) did not sit the Junior Certificate or (ii) sat the Junior Certificate in a year other than 1999 or 2000 (PISA 2000 sample) or 2002 or 2003 (PISA 2003 sample). A cross-tabulation of the special needs and left-school groups for PISA 2000 indicates that 23.3% of students with special needs were in first or second year at the time of the assessment and hence would be unlikely to have sat the Junior Certificate in 1999 or 2000. The percentage of students that left school in first and second year at the time of the assessment is also

quite high at 11.8%. Across all sampled students, just 4.6% were in first or second year at the time of the assessment. Similarly, in PISA 2003, 30.2% of special needs students were in first or second year at the time of the PISA assessment, and 55.2% of the left-school group was in first or second year.

Table 3.7. Mean Junior Certificate Mathematics Performance Scores (MJCPS) for the PISA 2003 Cohort, by Participation Status

Group	N total	N available	% available	MJCPS Mean	SD	SE
Present	3880	3645	93.9	8.41	2.11	.034
Absent	761	671	88.2	7.31	2.36	.091
Refusal	91	78	85.7	7.55	2.25	.255
Special Needs	125	81	64.8	5.46	1.77	.196
Left School	139	61	43.9	5.95	2.58	.331
Total	4996	4536	90.8	8.15	2.25	.033

Note. Estimates are unweighted. The standard errors reported in this table are somewhat underestimated since the clustered nature of the sample design has not been taken into account in the analysis. The total N does not include 4 age-ineligible students.

In general, the means for Junior Certificate mathematics are lower than the means for English, presumably a reflection of differences in the proportions of students taking the examination in these two subjects at higher level (close to two-thirds of students take English at Higher level, compared to just two-fifths who take mathematics at that level; see Shiel et al., 2001, pp. 141-144; Cosgrove et al., 2005, pp. 168-170). The mean EJCPS score of the group who participated in PISA 2000 (9.19) is the highest of the five groups in that year. Similarly, the mean MJCPS score of participating students in PISA 2003 (8.41) is the highest of the five groups.

Two one-way ANOVAs comparing mean scores across the five groups for 2000 and 2003 reveal that the differences are highly significant in both 2000 [$F(4, 4339) = 82.027$; $p < .001$] and 2003 [$F(4, 4531) = 87.216$; $p < .001$].

Post-hoc comparisons of individual means adjusted using both the Bonferroni and the Scheffé methods (the most conservative comparison methods[31]) indicate, for both PISA 2000 EJCPS scores and PISA 2003 MJCPS scores, that

[31] A comparison of the confidence intervals and significance levels produced by the two *post-hoc* comparison methods reveal no differences to 10 decimal places.

(i) the group which was present scored significantly higher than all other groups;

(ii) the absent and refusing groups do not differ; both groups scored significantly lower than the present group, and significantly higher than the special needs and left-school groups; and

(iii) the left-school and special needs groups do not differ from each other, both scoring significantly lower than the other three groups.

Although the increase in the standard error of measurement associated with the clustered sample design has not been controlled for in this analysis, all statistically significant differences between the pairwise comparisons have an associated p-value of .00002 or less. It is highly likely that the observed differences would be significant if the clustered design had been controlled for.

3.5.4. Logistic Regression of Student Participation Status

Since the aim of the analysis is to compare students who participated in PISA with those who *should have* but did not, and since the refusing and absent students do not differ in ability in terms of performance on the Junior Certificate, the analysis proceeds with a set of binary logistic regressions, with did participate/did not participate as the outcome. That is, students who have left the school and who have special educational needs are excluded from the analysis, and the absent and refusing groups are combined, because they do not differ in performance on EJCPS (PISA 2000) or MJCPS (PISA 2003). The analysis is unweighted since both the PISA school and student weights include a non-response adjustment.

3.5.4.1. Method

Variables. The following were used as explanatory variables for both PISA 2000 and PISA 2003:

Student-level variables

- student sex (reference category = male)
- student grade level (interval, second year to fifth year; reference category = third year)[32]

[32] Three students were in first year and one was in sixth year in PISA 2000; while in PISA 2003, coincidentally, three students were in first year and one in sixth year. In these eight cases, grade was recoded to the next nearest value to avoid very small cell counts.

- level at which Junior Certificate English/mathematics examination was taken (categories: higher, ordinary, foundation; reference category = ordinary)

School-level Variables

- school designated disadvantaged status (categories: yes, no; reference category = yes)
- school sex composition (categories: single sex, mixed sex; reference category = single sex)
- school type (categories: secondary, community/comprehensive, vocational; reference category = secondary).

3.5.4.2. Procedure

Following model-building procedures similar to those used for analyses of school participation, each explanatory variable was first tested separately and its significance evaluated using the Wald chi-square statistic. Next, all student-level variables were entered simultaneously and non-significant variables were removed in sequence using a backwards elimination strategy. All school-level variables were then added to the model, and non-significant variables removed in sequence, again using a backwards elimination strategy. Finally, tests for interactions were carried out before finalising the model.

3.5.4.3. Results

Table 3.8 shows the outcomes for each variable tested separately for PISA 2000. Boys and girls are about equally likely to participate. Students in third and fifth year are also equally likely to participate but students in fourth year and particularly second year are less likely to participate compared with third year students (in both cases, $p < .001$). While those taking the Junior Certificate English examination at higher level are more likely to participate than those taking the examination at ordinary level ($p < .001$), those taking foundation level are somewhat less likely ($p = .079$; borderline significant). Students in schools designated as disadvantaged are less likely than those in non-designated schools to participate ($p < .001$). While students in community/ comprehensive schools are about as likely as students in secondary schools to participate ($p = .892$), those in vocational schools are less likely ($p < .001$). Similarly, students in mixed-sex schools are less likely to participate than those in single-sex schools ($p < .001$).

Table 3.8. Binary Logistic Regressions with Student Participation as the Outcome Variable, and Six Explanatory Variables Tested Separately: PISA 2000

Variable/Comparison	*B*	*SE (B)*	*df*	*p*	*Exp (B)*
Student Sex					
Female - Male	-0.081	0.082	1	.326	0.922
Student Year Level					
Second year - Third Year	-0.944	.179	1	<.001	0.389
Fourth Year - Third Year	-0.741	.101	1	<.001	0.477
Fifth Year - Third Year	-0.086	.115	1	.451	0.917
Student Level of English					
Higher - Ordinary	1.043	.084	1	<.001	2.836
Foundation - Ordinary	-0.395	.225	1	.079	0.673
Designated Status of School					
Designated - Not Designated	-0.451	0.088	1	<.001	0.637
School Type					
Community/Comprehensive - Secondary	-0.018	.129	1	.892	0.983
Vocational - Secondary	-0.625	.092	1	<.001	0.535
School Sex Composition					
Mixed Sex - Single Sex	-0.364	0.086	1	<.001	0.695

Note. Data are unweighted. Standard errors do not take the clustered nature of the sample design into account.

In PISA 2003, boys and girls are once again about equally likely to participate. Students in fifth year are somewhat less likely to participate than students in third year ($p = .074$, borderline significant), but students in fourth year and, even more so, second year, are less likely to participate compared with third year students (in both cases, $p < .001$). Students taking the Junior Certificate mathematics examination at higher level are more likely to participate than those taking the examination at ordinary level, while those taking foundation level are less likely (in both cases, $p < .001$). Students in schools designated as disadvantaged are less likely than those in non-designated schools to participate ($p < .001$). While students in community/comprehensive schools are somewhat less likely as students in secondary schools to participate ($p = .084$, borderline significant), those in vocational schools are significantly less likely ($p < .001$). Students in mixed-sex schools are less likely to participate than those in single-sex schools ($p < .001$) (Table 3.9).

Table 3.9. Binary Logistic Regressions with Student Participation as the Outcome Variable, and Six Explanatory Variables Tested Separately: PISA 2003

Variable/Comparison	*B*	*SE (B)*	*df*	*p*	*Exp (B)*
Student Sex					
Female - Male	-070	.076	1	.357	0.933
Student Year Level					
Second year - Third Year	-1.055	.171	1	<.001	0.348
Fourth Year - Third Year	-.562	.096	1	<.001	0.570
Fifth Year - Third Year	-.181	.101	1	.074	0.835
Student Level of Mathematics					
Higher - Ordinary	.515	.092	1	<.001	0.360
Foundation - Ordinary	-1.023	.121	1	<.001	0.598
Designated Status of School					
Designated - Not Designated	-.407	.081	1	<.001	0.665
School Type					
Community/Comprehensive - Secondary	.188	.109	1	.084	1.207
Vocational - Secondary	-.454	.087	1	<.001	0.635
School Sex Composition					
Mixed Sex - Single Sex	-.292	.078	1	<.001	0.747

Note. Data are unweighted. Standard errors do not take the clustered nature of the sample design into account.

For each year, the significant student-level variables were then entered simultaneously. In the case of PISA 2000, both variables remain significant in the presence of one another; the coefficients for year level are similar to when tested separately; and the coefficient for the comparison of foundation and ordinary level is now statistically significant (p = .047) compared to borderline significant (p = .079) when tested separately (Table 3.10).

Table 3.10. Binary Logistic Regressions with Student Participation as the Outcome Variable, and Two Student-Level Variables Tested Together: PISA 2000

Variable/Comparison	*B*	*SE (B)*	*df*	*p*	*Exp (B)*
Student Year Level					
Second year - Third Year	-0.625	0.185	1	.001	0.535
Fourth Year - Third Year	-0.890	0.105	1	<.001	0.410
Fifth Year - Third Year	-0.121	0.117	1	.301	0.886
Student Level of English					
Higher - Ordinary	1.078	0.086	1	<.001	2.939
Foundation - Ordinary	-0.440	0.222	1	.047	0.644

Note. Data are unweighted. Standard errors do not take the clustered nature of the sample design into account.

In the case of PISA 2003, both variables once again remain significant in the presence of one another. The coefficients for syllabus level remain similar, but the results for year level differ. First, when tested together with syllabus level, second years are about as

likely as third years to participate; but the small numbers of second years result in a large standard error associated with the coefficient. Second, the difference in participation likelihood between fifth and third years is now statistically significant (p = .029), rather than borderline significant (p = .074) when tested separately (Table 3.11).

Table 3.11. Binary Logistic Regressions with Student Participation as the Outcome Variable, and Two Student-Level Variables Tested Together: PISA 2003

Variable/Comparison	*B*	*SE (B)*	*df*	*p*	*Exp (B)*
Student Year Level					
Second year - Third Year	-.759	1.162	1	.514	0.468
Fourth Year - Third Year	-.819	.102	1	<.001	0.441
Fifth Year - Third Year	-.233	.107	1	.029	0.793
Student Level of Mathematics					
Higher - Ordinary	.578	.093	1	<.001	1.782
Foundation - Ordinary	-1.146	.123	1	<.001	0.318

Note. Data are unweighted. Standard errors do not take the clustered nature of the sample design into account.

All school-level variables were then entered simultaneously with the two student-level variables. In the case of PISA 2000, the significance of the school-level variables is attenuated by the presence of the student-level variables (Table 3.12). For example, when designated status is tested on its own it can be seen that students from non-designated schools are about 1.7 times more likely than students in designated schools to participate, but when student year-level and Junior Certificate syllabus-level are taken into account, this reduces to about 1.2 times as likely, and just borderline significant (p = .066). The significance level of the vocational-secondary comparison decreases from when tested separately (p < .001) to when tested with the other school- and student-level variables (p = .039). School sex composition is no longer significant (p = .134). The effects of the student-level variables on participation likelihood remain largely similar; however the difference in participation likelihood of foundation level students compared to ordinary level is no longer significant (p = .104).

Table 3.12. Binary Logistic Regressions with Student Participation as the Outcome Variable, and Two Student-Level and Three School-Level Variables Tested Together: PISA 2000

Variable/Comparison	*B*	*SE (B)*	*df*	*p*	*Exp (B)*
Student Year Level					
Second year - Third Year	-0.622	0.186	1	.001	0.537
Fourth Year - Third Year	-0.913	0.106	1	<.001	0.401
Fifth Year - Third Year	-0.082	0.118	1	.486	0.921
Student Level of English					
Higher - Ordinary	0.969	0.089	1	<.001	2.636
Foundation - Ordinary	-0.355	0.219	1	.104	0.701
Designated Status of School					
Designated - Not Designated	-0.180	0.098	1	.066	0.835
School Type					
Community/Comprehensive - Secondary	0.169	0.150	1	.258	1.184
Vocational - Secondary	-0.246	0.120	1	.039	0.782
School Sex Composition					
Mixed Sex - Single Sex	-0.166	0.111	1	.134	0.847

Note. Data are unweighted. Standard errors do not take the clustered nature of the sample design into account.

In the case of PISA 2003, the significance of the school-level variables is not attenuated to the same degree as in PISA 2000 by the presence of the student-level variables. Designated disadvantaged status retains its statistical significance ($p = .003$), as does school sex composition ($p = .011$). The most striking difference occurs for school type. When tested separately, the participation likelihood of students in community/comprehensive and secondary schools was borderline significant. When tests with the other school and student variables, it becomes highly significant ($p < .001$), and the direction of the coefficient suggests that, after adjusting for the other variables in the model, students in community/comprehensive schools are more likely to participate than those in secondary schools. The coefficient for students in vocational schools compared with those in secondary schools is not significant in the presence of the other variables ($p = .947$) (Table 3.13).

Table 3.13. Binary Logistic Regressions with Student Participation as the Outcome Variable, and Two Student-Level and Three School-Level Variables Tested Together and Final Model: PISA 2003

Variable/Comparison	*B*	*SE (B)*	*df*	*p*	*Exp (B)*
Student Year Level					
Second year - Third Year	-.626	1.167	1	.592	.535
Fourth Year - Third Year	-.865	.104	1	<.001	.421
Fifth Year - Third Year	-.205	.107	1	.055	.814
Student Level of English					
Higher - Ordinary	.540	.094	1	<.001	1.716
Foundation - Ordinary	-1.046	.126	1	<.001	.351
Designated Status of School					
Not Designated - Designated	-.276	.094	1	.003	.759
School Type					
Community/Comprehensive - Secondary	.549	.141	1	<.001	1.731
Vocational - Secondary	-.014	.125	1	.911	.986
School Sex Composition					
Mixed Sex - Single Sex	-.290	.114	1	.011	.749

Note. Data are unweighted. Standard errors do not take the clustered nature of the sample design into account.

3.5.4.4. Final Models of Student Participation in PISA 2000 and PISA 2003

To finalise the models, a backwards elimination strategy was employed. For the PISA 2000 model, following the removal of school sex composition, all other variables remain significant, apart from designated disadvantaged status, which is borderline significant ($p = .066$). Following the approach taken in the development of multilevel models in the PISA 2000 national report for Ireland (Shiel et al., 2001, p. 98), a comparison of the difference between -2 log likelihood values (measures of overall fit) for the two models referred to a chi-square distribution with one degree of freedom (the difference in the number of parameters between the two models) shows that the inclusion of designated status does not significantly improve model fit ($\chi^2 = 3.395$; $df = 1$; $p = .065$) and so it is dropped from the model.

Then, tests for two-way interactions, within student and school levels, were carried out. For PISA 2000, there is no significant interaction between grade level and syllabus level; there is only one school-level variable so no test of interactions at the level of the school were necessary. Thus the final model for participation in PISA 2000 contains just two student-level and one school-level variable and no interaction terms (Table 3.14). The model suggests that students from vocational schools, those in second and fourth year, and those taking the Junior Certificate English examination at ordinary or foundation level, were less likely than other students to participate, while those in

community/comprehensive, secondary schools, those in third year and fifth year, and those taking English at higher level for the Junior Certificate, were more likely to have participated.

Table 3.14. Binary Logistic Regression with Student Participation as the Outcome Variable, and Two Student-Level Variables and One School-Level Variable Tested Together: Final Model – PISA 2000

Variable/Comparison	*B*	*SE (B)*	*df*	*p*	*Exp (B)*
Student Year Level					
Second year - Third Year	-0.625	0.185	1	.001	0.535
Fourth Year - Third Year	-0.898	0.106	1	<.001	0.407
Fifth Year - Third Year	-0.103	0.117	1	.381	0.902
Student Level of English					
Higher - Ordinary	1.000	0.088	1	<.001	2.719
Foundation - Ordinary	-0.390	0.218	1	.073	0.677
School Type					
Community/Comprehensive - Secondary	-0.027	0.134	1	.838	1.028
Vocational - Secondary	-0.386	0.085	1	<.001	0.680

Note. Data are unweighted. Standard errors do not take the clustered nature of the sample design into account.

In the case of PISA 2003, the backwards elimination strategy was not applied, since all variables retain statistical significance when tested simultaneously. All two-way interactions within levels were once again tested. The student-level variables do not interact significantly with one another. There are no interactions among the school-level variables. The model for student participation in PISA 2003 is thus the one shown in Table 3.13. The model indicates that students in second year are about as likely to participate as third years, while students in fourth and fifth year are less likely. Students taking higher level mathematics are significantly more likely to participate than ordinary level students, while students at foundation level are significantly less likely. Students in non-designated and mixed-sex schools are more likely to participate. Finally, while there is no difference in the participation likelihood of students in secondary and vocational schools, those in community/comprehensive schools are significantly more likely to participate.

3.5.4.5. Participation Rates of Groups of Students for Variables in the Models

A cross-tabulation of each of the variables in the final models with participation status shows the extent to which the two groups differ (Tables 3.15 for PISA 2000 and Table 3.16 for PISA 2003).

Table 3.15. Cross-tabulation of Student Year level, Junior Certificate English Syllabus Level, and School Type with Student Participation Status: PISA 2000

	Absent	*Row %*	*Present*	*Row %*	*Total*
Year Level					
Second Year	48	27.3	128	72.7	176
Third Year	346	12.7	2372	87.3	2718
Fourth Year	187	23.4	611	76.6	798
Fifth Year	118	13.7	742	86.3	860
JC English Syllabus Level					
Higher	300	10.3	2613	89.7	2913
Ordinary	268	18.9	1152	81.1	1420
Foundation	27	32.9	55	67.1	82
School Type					
Secondary	384	13.3	2503	86.7	2887
Vocational	233	22.3	813	77.7	1046
Comm/Comp	84	13.5	538	86.5	622
Total	701	15.4	3854	84.6	4555

Note. Data are unweighted.

Table 3.16. Cross-tabulation of Student Year level, Junior Certificate Mathematics Syllabus Level, School Type, School Designated Status, and School Sex Composition, with Student Participation Status: PISA 2003

	Absent	*Row %*	*Present*	*Row %*	*Total*
Student Year Level					
Second Year	57	34.1	110	65.9	167
Third Year	425	15.2	2362	84.8	2787
Fourth Year	207	24.0	654	76.0	861
Fifth Year	163	17.8	754	82.2	917
JC Mathematics Syllabus Level					
Higher	211	11.6	1601	88.4	1812
Ordinary	401	18.1	1818	81.9	2219
Foundation	137	37.7	226	62.3	363
School Type					
Secondary	465	16.8	2302	83.2	2767
Vocational	261	24.0	825	76.0	1086
Comm/Comp	126	14.3	753	85.7	879
School Designated Status					
Designated	298	22.6	1021	77.4	1319
Non-Designated	554	16.2	2859	83.8	3413
School Sex Composition					
Single Sex	315	15.6	1705	84.4	2020
Mixed Sex	537	19.8	2175	80.2	2712
Total	852	18.	3880	82.0	4732

Note. Data are unweighted.

In PISA 2000, the under-representation of second and fourth year students, those taking Junior Certificate English at ordinary and foundation level, and those in vocational schools, is apparent, when compared with overall participation rates. In the case of PISA 2003, the under-representation of second and fourth years, of foundation level

students, and of students in vocational, designated disadvantaged and mixed sex schools is also apparent. However, it should be recalled that when school type and student syllabus level were entered together, the participation likelihood of students in vocational schools does not differ to those in secondary schools, and the participation likelihood of community/comprehensive schools is significantly higher. This may relate to the fact that different proportions of students took Junior Certificate mathematics at higher, ordinary and foundation levels. The percentages of students taking higher level mathematics in secondary, community/comprehensive and vocational schools, respectively, are 43.5%, 36.7% and 26.2%. Corresponding percentages for foundation level are 5.2%, 6.8% and 14.7%.

While the school-level weight includes adjustments for school non-response by school type and sex composition, under-representation of *students* in different types of school is not controlled for by the school-level non-response adjustment, since the models in Tables 3.13 and 3.14 examined *student*, not *school* non-response. Similarly, differential participation rates by syllabus level and year level (in both 2000 and 2003) and by school sex composition and designated status (in 2003) is not controlled for.

3.6. Comparison of Between-School Variance and Total Variance for All Available Junior Certificate Achievement Data with Data for Students Participating in PISA 2000 and PISA 2003

As noted in Chapter 2, a second possible consequence of non-response is a downward bias in both total variance and between-school variance. The aim of this section is therefore to explore potential biases in the total variance and the between-school variance in achievement that may have arisen as a result of student non-response. This is accomplished through a comparison of achievement variance (total and between-school) of the JCPS scales associated with the PISA 2000 and 2003 samples of participating students with the more 'complete' sample of student Junior Certificate Examination data.

Since they are on different scales, one cannot directly compare the achievement variance of PISA with that of the Junior Certificate Examinations. However, if a difference is found between JCPS scores of participating students and all students for

whom data are available, whether they participated in PISA or not, a similar difference is assumed to be the case for the PISA achievement data.

Total and between-school variance is compared for the three groups for both years (JCPS scores for PISA participants, and JCPS scores for all available cases). It is hypothesised that both the between-school variance and the total variance will be greater for the 'complete' sample since this includes more low achievers. Since the analyses in Chapter 5 include only those students who attempted the Junior Certificate in the same year as the PISA survey, I also compare the total and between-school variance of this subset with all students participating on PISA, to investigate whether the selection of this subset has a substantial impact on the variance components.

3.6.1. Procedure

To explore the two hypotheses regarding student non-response, I compare the total variance and between-school variance of the following groups:

- EJCPS scores of all students participating in PISA 2000 with EJCPS scores of *all* eligible students, both participating and non-participating, for whom these data are available for the 1999 and 2000 Junior Certificate Examinations.
- MJCPS scores of all students participating in PISA 2003 with MJCPS scores of *all* eligible students, both participating and non-participating, for whom these data are available for the 2002 and 2003 Junior Certificate Examinations.

I use the original 12-point JCPS scales (Table 3.5) in these analyses since the 'preferred' scales (described in Chapter 4) have been developed on the basis of the obtained PISA samples (i.e., students who participated in PISA). Collapsing the lower end of the scale may well disguise precisely the differences in variance that I expect to find in these comparisons. Analyses are carried out in HLM 6.0 (Raudenbush, Bryk, Cheong, & Congdon, 2004), without the use of sampling weights.

3.6.2. Achievement Variance and Student Non-Response: PISA 2000

Table 3.17 shows the total, between-, and within-school variance of (i) all students selected to participate in PISA 2000, who took the Junior Certificate Examination in 1999 or 2000 (N = 4348), and (ii) students who participated in PISA 2000 and who took the Junior Certificate Examination in 1999 or 2000 (N = 3625). Two observations may

be made from these data. First, the percentage of variance attributable to schools is similar for the two groups, in the region of 19% to 20%. Second, the total variance associated with all available students (3.32) is about 13% larger than the total variance of students who actually participated in PISA 2000 (2.94). Thus, in PISA 2000, there is no support for the hypothesis regarding between-school variance, but there is some support for the hypothesis regarding total variance. That is, student non-response appears to be unrelated to between-school variance, but student non-response may have resulted in an underestimation in the total variance in achievement.

Table 3.17. *Variance Components Associated With the EJCPS from the PISA 2000 Sample: All Available Students and All Participating Students*

	All Available (N = 4348)		*All Participating*	
	Variance	% of total	Variance	% of total
Between students	2.6719	80.4	2.3890	81.2
Between schools	0.6495	19.6	0.5532	18.8
Total variance	3.3214	100.0	2.9422	100.0

Note. Analyses are unweighted. Data exclude approximately 6% of students who took the examination in a year other than 1999 or 2000.

3.6.3. Achievement Variance and Student Non-Response: PISA 2003

Table 3.18 shows the total, between, and within school variance of (i) all students selected to participate in PISA 2003, who took the Junior Certificate Examination in 2002 or 2003 (N = 4394), and (ii) students who participated in PISA 2003 and who took the Junior Certificate Examination in 2002 or 2003 (N = 3644). As with Junior Certificate English, the percentage of variance attributable to schools is similar for the three groups, in the region of 20% or 21%. Second, the total variance associated with all available students (4.9) is about 7% larger than the total variance of students who participated in PISA 2000 (4.6). Therefore there is some support for the hypothesis regarding total variance, but no support for the hypothesis regarding between-school variance.

Table 3.18. *Variance Components Associated With the MJCPS from the PISA 2003 Sample: All Available Students and All Participating Students*

	All Available (N = 4394)		*All Participating*	
	Variance	% of total	Variance	% of total
Between students	3.8345	78.7	3.6518	80.0
Between schools	1.0348	21.3	0.9103	20.0
Total variance	4.8693	100.0	4.5621	100.0

Note. Analyses are unweighted. Data exclude approximately 6% of students who took the examination in a year other than 2002 or 2003.

3.7. Conclusion

The analyses presented in this chapter arose from concerns about PISA's standards for response rates, and the statistical adjustments used to account for non-response at the school and student levels. They aim to address the question of what PISA can tell us about the achievements of students in Ireland, particularly as it relates to the reliability of achievement estimates at the extremes of the performance distribution and certain subgroups of the population.

A series of logistic regressions at both the school and student levels was carried out, with participation status as the outcome variable, and various school and student characteristics as explanatory variables. In the case of PISA 2000, participating and non-participating schools are similar with respect to the variables examined. In PISA 2003, secondary schools were significantly less likely to participate, although the non-response adjustments applied to schools are made with respect to both explicit strata (school size) and implicit strata (school type and school sex composition). In both PISA 2000 and PISA 2003, the mean achievements of participating and non-participating schools do not differ. In both years, non-response bias at the student level is evident. Absent and refusing students have significantly lower achievement on the Junior Certificate (English in 2000 and mathematics in 2003) than students who participated. In PISA 2000, while male and female students were about equally likely to participate, students in vocational schools, in second and fourth year, and taking Junior Certificate English at ordinary and foundation levels were significantly under-represented. When school- and student-level variables are examined simultaneously, there is only a borderline significant difference in PISA 2000 in the student-level participation rates of schools designated disadvantaged and those not so designated. In the case of PISA 2003, second and fourth year students, and those taking Junior Certificate mathematics

at ordinary and foundation levels are also under-represented. In contrast to the model for PISA 2000 however, two variables in addition to school type – school designated status and school sex composition – were significantly predictive of student non-response. The differences in the models suggest that the expectations set up as a result of students being told by school staff that they had been selected to participate in a test of reading (in PISA 2000) or a test of mathematics (in PISA 2003) may have given rise for student absenteeism for somewhat different reasons in the two surveys.

As pointed out in Chapter 2, mean PISA scores differ across year levels in both 2000 and 2003. They also differ substantially across Junior Certificate English syllabus level (PISA 2000) and mathematics syllabus level (PISA 2003). Mean scores vary significantly across school type also in both PISA 2000 and PISA 2003. And, in the case of PISA 2003, where the final model of student participation also included school sex composition and school designated disadvantaged status, it should again be noted that significant achievement differences are associated with these two variables.

One could conclude from the analyses of student non-response in PISA 2000 and PISA 2003, together with other research evidence presented in Chapters 1 and 2, that the differential rates of student participation across school and student variables resulted in somewhat of an overestimation of the achievement of Irish students as measured by the PISA 2000 test of reading literacy and the PISA 2003 test of mathematics. Further, the analyses suggest that mean scores of some subgroups are not as reliable as others (e.g., foundation level students in both 2000 and 2003; students in vocational schools in 2003). The analyses also suggest that any conclusions drawn about reading and mathematics standards of lower-achieving students are particularly problematic since many students in this group were not present for the PISA assessment, despite the fact that they were eligible to participate.

In the PISA 2000 national report for Ireland, a Pearson correlation of .74 was reported between performance on PISA 2000 reading literacy and Junior Certificate English for students taking the examination in 1999 or 2000 (i.e., 94% of the students that participated in PISA 2000). In 2003, the correlation between performance on PISA mathematics and Junior Certificate mathematics was similar (.75). Thus, while the overall relationship between the two assessments is substantial, analyses of the

relationship between the two assessments at the lower end of the scale, and for some subgroups, will be less reliable than overall comparisons, or those made at the middle or upper portions.

The second set of analyses reported here examined the consequences of non-response at the student level in the interpretation of total and within-school variance. This was done by comparing variance on the Junior Certificate Examinations for all students, whether present or absent, with variance on the Junior Certificate Examinations for only students who participated in PISA. The analyses assume, therefore, that patterns of variance on PISA are similar to the Junior Certificate Examinations. Results indicated that between-school variance on Junior Certificate English (in 2000) or mathematics (in 2003) does not differ between the obtained PISA samples and the available Junior Certificate data for all sampled students, regardless of whether they participated or not. However, there is some evidence that student non-response results in a downward bias in the total variance, particularly in the case of Junior Certificate English. Since it is probable that the majority of the 'missing' variance pertains to the lower end of the achievement distribution, these findings provide further support for the argument that, given student non-response, reliable inferences regarding standards of lower-achieving students are not possible.

Overall, these analyses call into question the validity of published results which cite the percentage of students at or below proficiency Level 1 on the achievement scales. This statistic is widely cited as an indicator of the percentage of students in a country or group with low literacy levels. The OECD has cited Level 1 as a minimum requirement for successful participation in society. Level 2 in PISA 2003 mathematics is viewed as "a baseline level of mathematics proficiency on the PISA scale at which students begin to demonstrate the kind of literacy skills that enable them to actively use mathematics" (OECD, 2004c, p. 69). In PISA 2000, students at or below Level 1 on the reading proficiency scale were described as having ... "serious difficulties in using reading literacy as an effective tool to advance their knowledge and skills in other areas. Students with literacy skills below Level 1 may, therefore, be at risk not only of difficulties in their initial transition from education to work but also of failure to benefit from further education and learning opportunities through life" (OECD, 2001b, p. 48). The importance of the percentages of students at the lowest point of the proficiency

levels was also noted in Irish media reports and ministerial speeches discussed in Chapter 1. Further, the tentative evidence from a re-analysis of IALS linking growth in literacy levels at the lower end of the achievement distribution with economic growth (Coulombe et al., 2004) adds a further reason for the OECD to re-examine the current methods for adjustments of non-response at the student level.

In PISA, and in international studies preceding it, the same sampling standards were applied across participating education systems. This seems illogical, given that (i) there are considerable variations across countries in the proportion of variance within and between schools, and (ii) there are differences between countries regarding the strength of the relationship between rate of participation and achievement. Alternative methods for better adjusting for non-response (e.g., the use of imputation methods to produce achievement scores for students who did not participate) have been proposed. Monseur and Wu (2002) have explored imputation methods to control for non-response and have shown that these reduce the bias in estimates of means as well as in the variance in achievement. At the very least, it would seem timely to revisit the issue of sampling standards, and perhaps raise the minimum student response rate standard of PISA (80%), particularly in systems where there is a relationship between student response rate and achievement, coupled with a comparatively high within-school variance. In Ireland's case, where student response rates were lower than in most participating countries in both PISA 2000 and 2003, strategies to encourage higher student participation rates in future PISA cycles should be developed.

CHAPTER 4. ANALYSES OF ACHIEVEMENTS ON JUNIOR CERTIFICATE ENGLISH AND MATHEMATICS EXAMINATIONS USING ACHIEVEMENT DATA FROM PISA

4.1. Introduction

In Chapter 2, it was noted that the 12-point Junior Certificate Performance Scale (JCPS) has been used in a number of analyses, including comparisons of performance with PISA 2000 and PISA 2003, but that the possibility that achievement at the lower end of the scale may be more clustered than at the middle or upper ends has not been investigated, nor has the appropriateness of having an equal interval scale at the extremes. Further, comparisons between various versions of the JCPS and achievement on PISA 2003 have not yet been made. To address these issues, analyses described in this chapter explore various ways in which to scale the Junior Certificate English and mathematics achievement data and these will be used in analyses that will be reported in Chapter 5. It is hoped that the analyses may yield additional insights into what PISA can tell us about the achievements of students in Ireland when results are considered in the context of the literature review in Chapter 2.

4.2. Rationale

Comparisons of the content of PISA and the Junior Certificate reviewed in Chapter 2 yield substantial information on how the assessments differ from one another, as well as indicating areas of overlap. However, the extent to which performances on the two assessments are related has not been fully explored. In the case of mathematics in particular, where PISA and the Junior Certificate differ markedly, further analysis of how the assessments may relate to one another could be of benefit. To address the limitations of the exploratory analyses of various forms of the JCPS reported in Shiel et al. (2001), analyses presented here explore the consequences of rescaling the foundation-level grades in particular. Preference is given to a scale which produces roughly equal intervals between JCPS scale points (as indicated by mean score differences on the PISA scales). To capitalise on the information available to guide the choice of an alternative Junior Certificate scale for English (using PISA 2000 data), and one for mathematics (using data from PISA 2003), the mean scores of students are compared for the combined reading scale and each of the three process subscales (in the

case of PISA 2000), and for the combined mathematics scales and each of the four content area subscales (in the case of PISA 2003).

4.3. Method

The analyses use three approaches:

(i) *Visual exploratory analyses.* Distributions of achievement were examined via frequency distributions of mean achievement at each syllabus level. Frequency distributions were produced in SPSS using the average of the five achievement estimates (plausible values), rounded to the nearest multiple of 5. The overlaps between the distributions of the syllabus levels are also discussed in terms of percentile points. These were estimated in SPSS using the average of the five plausible values. Since measurement and sampling error are not taken into account in analyses in SPSS, standard errors associated with these estimates are not reported. In analyses, data were weighted using the standardised student weight, which adjusts for differential sampling fractions and non-response at both the school and student levels (i.e., adjusts for the fact that schools and students were not sampled with equal probability, and that some schools and students did not participate in the assessment).

(ii) *Comparison of mean scores* at each point on the various JCPS scale possibilities. Mean scores were computed in WesVar 2.0 and, using variance replication techniques (specifically, Fay's variant of Balanced Repeated Replication) (Westat, 2000), both measurement and sampling error were taken into account. Since several comparisons are being made simultaneously, one should consider adjusting the confidence intervals to a more conservative level. For example, one could adjust the 95% confidence interval using Bonferroni's procedure (which entails dividing the alpha level by the number of comparisons; in the case of the 12-point scale, this would be 11, resulting in an overall alpha level of .0045; Dunn, 1961). However, it has been argued that the Bonferroni procedure results in conservative estimates especially with larger numbers of comparisons (see Cosgrove et al., 2005, p. 51). Further, it is only of interest here to compare adjacent groups (e.g., comparing a group with the one immediately below and/or above it). Applying the Bonferroni adjustment to the present analyses, one would run the risk of identifying too few groups (i.e., 'over-collapsing' letter grades). For this reason, the 95% confidence intervals, computed using WesVar, are not adjusted.

The comparisons of mean scores were made in four phases. In the first phase, different ways of collapsing the foundation level letter grades were explored. In the second, different overlaps between the syllabus levels were explored. However, it was expected, given the analyses of the 10-,12-, and 14-point scales reported in Shiel et al. (2001), that a three-grade overlap would prove to be the optimal way to scale the Junior Certificate English examination data. In the third phase, the mean score differences of adjacent groups was examined to see whether stretching the scale, particularly at the extremes, might give a 'smoother' scale. Finally, to account for the possibility that the choice of an alternative way to scale the JCPS may vary across PISA cycles due to sampling fluctuations or other variations, the mean combined reading scores of students at each point of the 12-point EJCPS were computed both for all students participating in PISA 2003 and also for the subset of students participating in PISA 2003 who took the Junior Certificate in 2003; similarly, the mean combined mathematics scores of students at each point of the 12-point MJCPS were computed for the PISA 2000 cohort and sub-cohort (i.e., students attempting the Junior Certificate in 2000). These were then compared to the equivalent means for the other PISA cycle in question.

(iii)*Pearson correlations* between each version of the JCPS and achievement on PISA (for both the overall scales and subscales; English in the case of PISA 2000 and mathematics in the case of PISA 2003). These overall measures of association were computed as a final check to confirm the appropriateness of the choice of the JCPS. They were obtained in WesVar using the regression function, and, for each set of regressions computed (i.e., each had to be computed five times, once with each plausible value), the correlation coefficient was calculated using a template in Excel which combines and transforms the regression coefficients into a Pearson correlation coefficient, taking both the lack of asymptotic Normal distribution and between-imputation variance into account (for a specific description, see Shiel et al., 2001, pp. 205-206).

4.4. Results: Junior Certificate Performance Scale for English (EJCPS)

4.4.1. Visual Exploratory Analyses

Figure 4.1 shows the distribution of PISA 2000 combined reading scores by syllabus level, expressed both as counts and percentages. The distribution of percentages at foundation level is more uneven than distributions at higher and ordinary levels. Higher

and ordinary levels have percentage distributions that are negatively skewed (skewness = -.226 and -.180, respectively), while foundation level is positively skewed (skewness = .385). The mean of the higher level group (562.1; SE = 2.12) is around the 62nd percentile of the overall achievement distribution; that of the ordinary level (450.9; SE = 3.89) at the 20th, and foundation level (336.0; SE = 9.80) at the 3rd.[33]

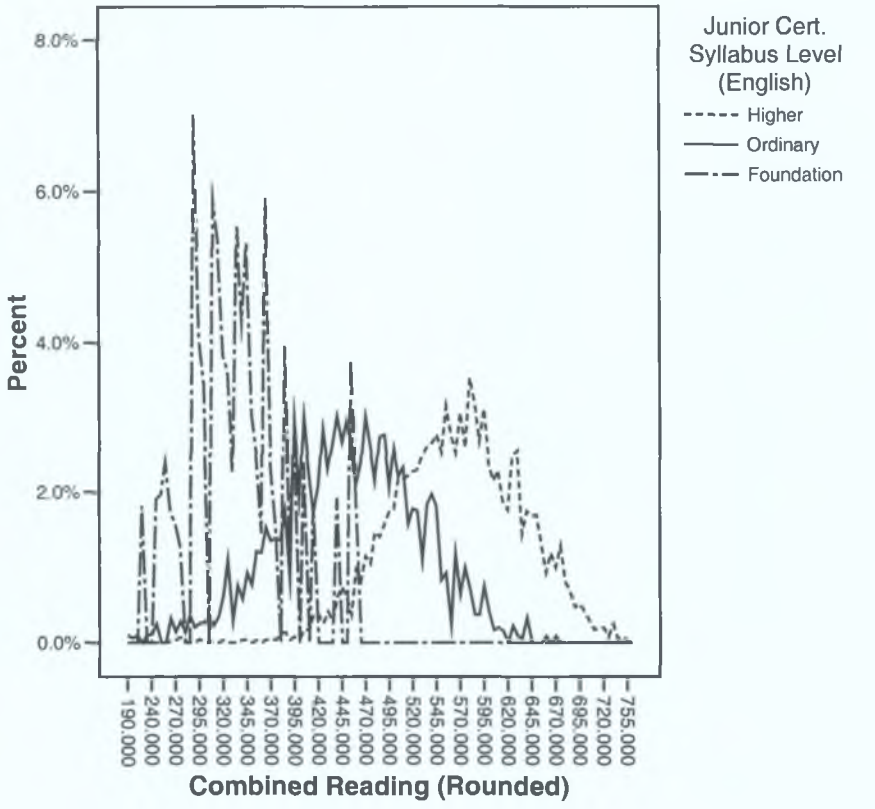
There is more of an overlap between the distributions at higher and ordinary levels, than between ordinary and foundation. Further, there is little overlap between higher and foundation levels. The intersection between the higher and ordinary level achievement distributions is around 520 score points which corresponds to the 25th percentile of the higher-level distribution, and the 82nd percentile at ordinary level. The intersection between the ordinary and foundation level achievement distributions is around 380 score points which corresponds to the 16th percentile of the ordinary level distribution, and the 83rd percentile at foundation level. The achievement distributions for higher and foundation levels overlap at around 470 score points, which corresponds to the 99th percentile at foundation level, and between the 7th and 8th percentile at higher.

4.4.2. Exploration of Mean Reading Scores Associated With Various Versions of the EJCPS: PISA 2000 Cohort

Table 4.1 shows the eight Junior Certificate English scales (EJCPS) explored using the PISA 2000 reading achievement data. For each of the eight scales, the combined reading literacy scores, as well as scores on the three reading process subscales are compared, using 95% confidence intervals to ascertain the extent of overlap between each point on the EJCPS.

[33] The percentile points and scale scores referred to in this section are approximate are based on the mean of the five plausible values and computed in SPSS. Hence, although the point estimates are accurate, the errors around the estimates would not be accurate using this method and so are not provided; this exploration is descriptive rather than explanatory in any case. The exceptions to this are the mean score estimates for Higher, Ordinary and foundation levels, which are taken from Shiel et al. (2001, Table 4.28) and were computed in WesVar, taking both sampling and measurement error into account.

Figure 4.1. *Frequency Distribution of PISA 2000 Combined Reading Scores by Higher, Ordinary and foundation Level English by (a) Percent and (b) Number (Count) of Cases Within Syllabus Level*



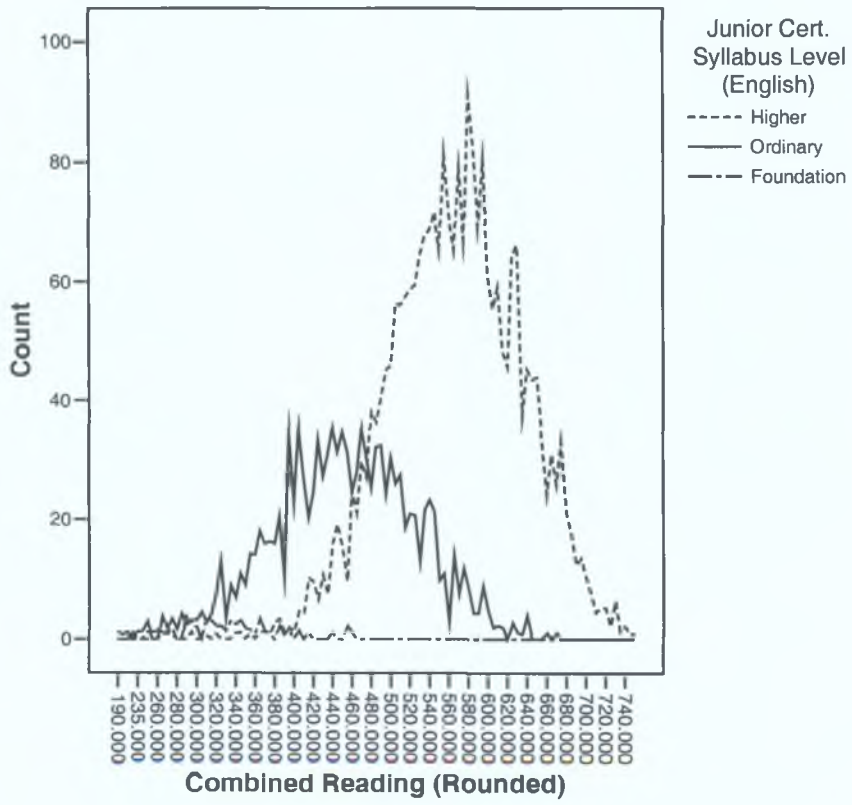


Table 4.1. Description of EJCPS Scoring Schemes Explored

Scale 1	*Grade/Level*		
Scale point	*Higher*	*Ordinary*	*Foundation*
12	A		
11	B		
10	C		
9	D	A	
8	E	B	
7	F	C	
6		D	A
5		E	B
4		F	C
3			D
2			E
1			F

Scale 5	*Grade/Level*		
Scale point	*Higher*	*Ordinary*	*Foundation*
11	A		
10	B		
9	C		
8	D		
7	E		
6	F		
5		A	
4		B	
3		C	
2		D	A
1		E, F	B, C, D, E, F

Scale 2	*Grade/Level*		
Scale point	*Higher*	*Ordinary*	*Foundation*
9	A		
8	B		
7	C		
6	D	A	
5	E	B	
4	F	C	
3		D	A
2		E	B
1		F	C, D, E, F

Scale 6	*Grade/Level*		
Scale point	*Higher*	*Ordinary*	*Foundation*
10	A		
9	B		
8	C		
7	D		
6	E		
5	F	A	
4		B	
3		C	
2		D	A
1		E, F	B, C, D, E, F

Scale 3	*Grade/Level*		
Scale point	*Higher*	*Ordinary*	*Foundation*
8	A		
7	B		
6	C		
5	D	A	
4	E	B	
3	F	C	
2		D	A
1		E, F	B, C, D, E, F

Scale 7	*Grade/Level*		
Scale point	*Higher*	*Ordinary*	*Foundation*
9	A		
8	B		
7	C		
6	D		
5	E	A	
4	F	B	
3		C	
2		D	A
1		E, F	B, C, D, E, F

Scale 4	*Grade/Level*		
Scale point	*Higher*	*Ordinary*	*Foundation*
8	A		
7	B		
6	C		
5	D	A	
4	E	B	
3	F	C	
2		D, E	A, B
1		F	C, D, E, F

Scale 8 - Final Scale	*Grade/Level*		
Scale point	*Higher*	*Ordinary*	*Foundation*
9	A		
8	B		
7	C		
6	D	A	
5	E	B	
4	F	C	
3		D	A
1		E, F	B, C, D, E, F

Description of scales

1 Original 12-point scale

2 9-point scale with original points 4, 3, 2, 1 collapsed into one category

3 8-point scale with original points 5, 4, 3, 2, 1 collapsed into one category

4 8-point scale with original points 4, 3, 2, 1 collapsed into one category and original 5 recoded to 6

5 11-point scale with no overlap between Higher and Ordinary level grades

6 10-point scale with a one-point overlap between Higher and Ordinary level grades

7 9-point scale with a two-point overlap between Higher and Ordinary level grades

8 8-point scale with original points 5, 4, 3, 2, 1 collapsed into one category and two scale points separating the lowest and second lowest categories

Table 4.2 shows the mean scores of all students participating in PISA 2000 for the combined reading literacy scale, and on the Retrieve, Interpret and Reflect subscales at each point on the original 12-point EJCPS. The table shows first of all, that only 10 students – well under 1% of the sample – are at the lowest three points on the scale, and that no student is at the lowest point.

Second, using the 95% confidence intervals as a guide, one cannot distinguish between the PISA scores of students at EJCPS scale points 2, 3, and 4. This relates both to the size of the standard errors that one would expect from such small groups of students and also to the clustering occurring around PISA mean scores. Only 62 or so scale points (for combined reading literacy, for example) separate the mean scores of students with an EJCPS score of 2 from those with a score of 5. In contrast, the mean PISA score difference at the upper end of the scale (between students scoring a 9 on the EJCPS and those scoring a 12) is about twice that for the lower end (around 116 scale points). Apart from the lowest four points on the scale, all points are empirically distinguishable (using the 95% confidence intervals as a guide).

Third, apart from the lowest four categories, there are ample numbers of students (a minimum of 167) at each point on the scale, and the associated standard errors are notably smaller for these groups.

Fourth, there is a 'clumping' of students scoring a 10 on EJCPS (28.1%). Fifth, there is an indication that students at the bottom of the EJCPS distribution did somewhat better on the Reflect subscale than on the other two subscales; however, there are too few students at these levels to draw any strong inferences. Apart from that, there are no notable differences between the mean scores of EJCPS groups across the combined reading scale and the reading subscales.

Table 4.2. *Means, Standard Errors, and 95% Confidence Intervals For the Combined Reading Score, and for the Three Reading Process Subscales, For Each Point on the EJCPS: Scale 1*

EJCPS	*N*	*%*	*Overall*	*SE*	*CI95L*	*CI95U*	*Retrieve*	*SE*	*CI95L*	*CI95U*	*Interpret*	*SE*	*CI95L*	*CI95U*	*Reflect*	*SE*	*CI95L*	*CI95U*
1*	0	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
2	3	0.1	**299.4**	25.33	249.0	349.8	**309.0**	29.23	250.8	367.1	**328.4**	27.36	274.0	382.9	**340.0**	34.47	271.4	408.6
3	7	0.2	**305.4**	23.76	258.1	352.7	**292.2**	36.92	218.8	365.7	**300.1**	30.77	238.9	361.4	**339.5**	23.20	293.4	385.7
4	27	0.7	**332.3**	10.39	311.6	352.9	**321.6**	13.82	294.1	349.1	**331.0**	13.55	304.0	357.9	**345.2**	13.00	319.3	371.1
5	23	0.6	**361.1**	19.43	322.5	399.8	**351.8**	18.80	314.4	389.2	**346.4**	18.88	308.8	384.0	**374.2**	18.02	338.4	410.1
6	208	5.4	**411.8**	5.85	400.2	423.5	**406.8**	6.63	393.6	420.0	**411.2**	7.232	396.8	425.6	**423.2**	6.98	409.3	437.1
7	505	13.1	**450.4**	4.38	441.6	459.1	**444.7**	4.70	435.4	454.1	**448.8**	4.668	439.5	458.1	**461.6**	4.55	452.5	470.6
8	339	8.8	**484.8**	4.15	476.5	493.0	**482.7**	4.26	474.2	491.1	**483.2**	4.413	474.4	491.9	**495.3**	3.92	487.5	503.1
9	638	16.5	**526.2**	3.88	518.5	533.9	**525.3**	4.09	517.2	533.4	**524.9**	4.067	516.8	532.9	**531.1**	4.42	522.3	539.9
10	1083	28.1	**561.6**	2.34	556.9	566.3	**561.1**	2.52	556.1	566.1	**563.0**	2.519	557.9	568.0	**567.0**	2.53	562.0	572.1
11	617	16.0	**602.6**	2.77	597.1	608.1	**600.5**	3.45	593.7	607.4	**603.0**	3.252	596.5	609.5	**604.3**	2.76	598.8	609.8
12	167	4.3	**642.4**	4.81	632.9	652.0	**639.1**	5.43	628.3	649.9	**644.9**	4.787	635.4	654.4	**642.8**	4.87	633.1	652.5
All available	3616	93.8	**531.1**	3.06	525.0	537.2	**528.8**	3.07	522.7	534.9	**530.9**	3.136	524.7	537.2	**537.3**	2.93	531.5	543.2
Missing	238	6.2	**459.7**	11.14	437.5	481.8	**455.8**	11.21	433.5	478.1	**459.0**	10.76	437.6	480.4	**469.7**	10.49	448.9	490.6
Total	3854	100.0	**526.7**	3.24	520.2	533.1	**524.3**	3.25	517.8	530.8	**526.5**	3.287	519.9	533.0	**533.2**	3.10	527.0	539.3

*In the PISA 2000 cohort, no student was awarded Grade F at Foundation Level

Table 4.3 shows mean PISA reading scores for the second EJCPS – i.e., a 9-point scale that results from collapsing EJCPS score categories 2, 3, and 4. The scale is now a little 'smoother', with a PISA combined reading score difference of between 35 and 40 scale points at each EJCPS point. However, an examination of the 95% confidence intervals suggests that the mean score differences of students in the bottom two EJCPS groups are not significantly different. Thus, scales 3 and 4 explore two alternative methods of collapsing the second-lowest group: collapsing the second lowest group with the lowest (scale 3) and collapsing the second-lowest group with the third-lowest (scale 4).

Table 4.4 shows the mean scores for the combined reading scale and three process subscales for scale 3. The 8-point scale now has empirically distinct mean scores (using the 95% confidence intervals as a guide). However, the mean score difference between the lowest and second lowest groups, ranging between about 68 points (Reflect subscale) and about 78 points (Interpret subscale), is somewhat larger than mean score differences between adjacent groups.

Table 4.5 shows the mean scores for scale 4. Collapsing the second lowest EJCPS group upwards rather than downwards results in an even larger gap between the new lowest and second lowest groups of between about 75 points (Reflect subscale) and about 86 points (Retrieve subscale). Further, the number of students in the lowest group is notably smaller than the number in the other groups. It was decided, therefore, to proceed with scale 3 and explore various ways of overlapping the EJCPS scale at higher and ordinary level.

Table 4.3. Means, Standard Errors, and 95% Confidence Intervals For the Combined Reading Score, and for the Three Reading Process Subscales, For Each Point on the EJCPS: Scale 2

EJCPS	N	%	Overall	SE	CI95L	CI95U	Retrieve	SE	CI95L	CI95U	Interpret	SE	CI95L	CI95U	Reflect	SE	CI95L	CI95U
1	36	1.0	**324.6**	9.31	306.0	343.1	**314.9**	11.94	291.1	338.7	**324.8**	12.47	299.9	349.6	**343.7**	10.90	322.0	365.4
2	23	0.6	**361.1**	19.43	322.5	399.8	**351.8**	18.80	314.4	389.2	**346.4**	18.88	308.8	384.0	**374.2**	18.02	338.4	410.1
3	208	5.8	**411.8**	5.85	400.2	423.5	**406.8**	6.63	393.6	420.0	**411.2**	7.232	396.8	425.6	**423.2**	6.98	409.3	437.1
4	505	14.0	**450.4**	4.38	441.6	459.1	**444.7**	4.70	435.4	454.1	**448.8**	4.668	439.5	458.1	**461.6**	4.55	452.5	470.6
5	339	9.4	**484.8**	4.15	476.5	493.0	**482.7**	4.26	474.2	491.1	**483.2**	4.413	474.4	491.9	**495.3**	3.92	487.5	503.1
6	638	17.6	**526.2**	3.88	518.5	533.9	**525.3**	4.09	517.2	533.4	**524.9**	4.067	516.8	532.9	**531.1**	4.42	522.3	539.9
7	1083	29.9	**561.6**	2.34	556.9	566.3	**561.1**	2.52	556.1	566.1	**563.0**	2.519	557.9	568.0	**567.0**	2.53	562.0	572.1
8	617	17.1	**602.6**	2.77	597.1	608.1	**600.5**	3.45	593.7	607.4	**603.0**	3.252	596.5	609.5	**604.3**	2.76	598.8	609.8
9	167	4.6	**642.4**	4.81	632.9	652.0	**639.1**	5.43	628.3	649.9	**644.9**	4.787	635.4	654.4	**642.8**	4.87	633.1	652.5
All available	3616	93.8	**531.1**	3.06	525.0	537.2	**528.8**	3.07	522.7	534.9	**530.9**	3.136	524.7	537.2	**537.3**	2.93	531.5	543.2
Missing	238	6.2	**459.7**	11.14	437.5	481.8	**455.8**	11.21	433.5	478.1	**459.0**	10.76	437.6	480.4	**469.7**	10.49	448.9	490.6
Total	3854	100.0	**526.7**	3.24	520.2	533.1	**524.3**	3.25	517.8	530.8	**526.5**	3.287	519.9	533.0	**533.2**	3.10	527.0	539.3

Table 4.4. Means, Standard Errors, and 95% Confidence Intervals For the Combined Reading Score, and for the Three Reading Process Subscales, For Each Point on the EJCPS: Scale 3

EJCPS	N	%	Overall	SE	CI95L	CI95U	Retrieve	SE	CI95L	CI95U	Interpret	SE	CI95L	CI95U	Reflect	SE	CI95L	CI95U
1	59	1.6	**338.6**	10.44	317.8	359.4	**329.1**	9.86	309.5	348.7	**333.1**	9.714	313.7	352.4	**355.4**	9.82	335.9	375.0
2	208	5.8	**411.8**	5.85	400.2	423.5	**406.8**	6.63	393.6	420.0	**411.2**	7.232	396.8	425.6	**423.2**	6.98	409.3	437.1
3	505	14.0	**450.4**	4.38	441.6	459.1	**444.7**	4.70	435.4	454.1	**448.8**	4.668	439.5	458.1	**461.6**	4.55	452.5	470.6
4	339	9.4	**484.8**	4.15	476.5	493.0	**482.7**	4.26	474.2	491.1	**483.2**	4.413	474.4	491.9	**495.3**	3.92	487.5	503.1
5	638	17.6	**526.2**	3.88	518.5	533.9	**525.3**	4.09	517.2	533.4	**524.9**	4.067	516.8	532.9	**531.1**	4.42	522.3	539.9
6	1083	29.9	**561.6**	2.34	556.9	566.3	**561.1**	2.52	556.1	566.1	**563.0**	2.519	557.9	568.0	**567.0**	2.53	562.0	572.1
7	617	17.1	**602.6**	2.77	597.1	608.1	**600.5**	3.45	593.7	607.4	**603.0**	3.252	596.5	609.5	**604.3**	2.76	598.8	609.8
8	167	4.6	**642.4**	4.81	632.9	652.0	**639.1**	5.43	628.3	649.9	**644.9**	4.787	635.4	654.4	**642.8**	4.87	633.1	652.5
All available	3616	93.8	**531.1**	3.06	525.0	537.2	**528.8**	3.07	522.7	534.9	**530.9**	3.136	524.7	537.2	**537.3**	2.93	531.5	543.2
Missing	238	6.2	**459.7**	11.14	437.5	481.8	**455.8**	11.21	433.5	478.1	**459.0**	10.76	437.6	480.4	**469.7**	10.49	448.9	490.6
Total	3854	100.0	**526.7**	3.24	520.2	533.1	**524.3**	3.25	517.8	530.8	**526.5**	3.287	519.9	533.0	**533.2**	3.10	527.0	539.3

Table 4.5. Means, Standard Errors, and 95% Confidence Intervals For the Combined Reading Score, and for the Three Reading Process Subscales, For Each Point on the EJCPS: Scale 4

EJCPS	*N*	*%*	*Overall*	*SE*	*CI95L*	*CI95U*	*Retrieve*	*SE*	*CI95L*	*CI95U*	*Interpret*	*SE*	*CI95L*	*CI95U*	*Reflect*	*SE*	*CI95L*	*CI95U*
1	36	1.0	**324.6**	9.31	306.0	343.1	**314.9**	11.94	291.1	338.7	**324.8**	12.47	299.9	349.6	**343.7**	10.90	322.0	365.4
2	231	6.4	**406.8**	5.62	395.7	418.0	**401.4**	6.27	388.9	413.9	**404.8**	6.768	391.4	418.3	**418.4**	6.58	405.3	431.5
3	505	14.0	**450.4**	4.38	441.6	459.1	**444.7**	4.70	435.4	454.1	**448.8**	4.668	439.5	458.1	**461.6**	4.55	452.5	470.6
4	339	9.4	**484.8**	4.15	476.5	493.0	**482.7**	4.26	474.2	491.1	**483.2**	4.413	474.4	491.9	**495.3**	3.92	487.5	503.1
5	638	17.6	**526.2**	3.88	518.5	533.9	**525.3**	4.09	517.2	533.4	**524.9**	4.067	516.8	532.9	**531.1**	4.42	522.3	539.9
6	1083	29.9	**561.6**	2.34	556.9	566.3	**561.1**	2.52	556.1	566.1	**563.0**	2.519	557.9	568.0	**567.0**	2.53	562.0	572.1
7	617	17.1	**602.6**	2.77	597.1	608.1	**600.5**	3.45	593.7	607.4	**603.0**	3.252	596.5	609.5	**604.3**	2.76	598.8	609.8
8	167	4.6	**642.4**	4.81	632.9	652.0	**639.1**	5.43	628.3	649.9	**644.9**	4.787	635.4	654.4	**642.8**	4.87	633.1	652.5
All available	3616	93.8	**531.1**	3.06	525.0	537.2	**528.8**	3.07	522.7	534.9	**530.9**	3.136	524.7	537.2	**537.3**	2.93	531.5	543.2
Missing	238	6.2	**459.7**	11.14	437.5	481.8	**455.8**	11.21	433.5	478.1	**459.0**	10.76	437.6	480.4	**469.7**	10.49	448.9	490.6
Total	3854	100.0	**526.7**	3.24	520.2	533.1	**524.3**	3.25	517.8	530.8	**526.5**	3.287	519.9	533.0	**533.2**	3.10	527.0	539.3

Scales 5, 6 and 7 explore the consequences of having no overlap, a one-grade overlap, and a two-grade overlap, respectively, between higher and ordinary levels. The PISA mean scores associated with each of these are shown in Tables 4.6, 4.7 and 4.8, respectively.

Scale 5 indicates that only one student obtained an E at higher level. Scales 5 and 6 suggest a dip in achievement on PISA around the middle of the EJCPS; that is, students obtaining grades E and F at higher level score lower than students obtaining grade A at ordinary level. A comparison of scale 7 and scale 3, which are identical except that scale 7 has a two-grade rather than a three-grade overlap between higher and ordinary, suggests that the original three-grade overlap may be more appropriate, given that the 95% confidence intervals for PISA mean scores associated with EJCPS score 9 and 10 overlap. However, the interval between the lowest and second lowest points on scale 7 is larger than the intervals between the other points on the scale. Therefore, scale 3 was re-scaled to range from 1 to 9, with a 2-point gap between the lowest and second lowest points of the scale. The average gap between scale points on the preferred scale (scale 8) is about 38 PISA scale points, or two-fifths of a (national) standard deviation.

4.4.3. Exploration of Pearson Correlations Associated With Various Versions of the EJCPS: PISA 2000 Cohort

Table 4.9 shows the Pearson correlations between the eight EJCPS scales explored and achievement on PISA 2000 combined reading and on the three reading process subscales. There is virtually no difference in the strength of correlation obtained, regardless of the EJCPS scale or the PISA scale considered. All correlations range between .69 and .72. There is a marginal decrease in the strength of the correlation as the overlap between higher and ordinary levels is decreased. The pattern of correlations supports the choice of scale 8, although the strength of its correlations with achievement are identical to three decimal places to scales 1 and 2. The advantage of scale 8 however, as noted, is that it is smoother and without overlap of the 95% confidence intervals between adjacent points. Figure 4.2 depicts scale 8 graphically, with the PISA combined mean reading score and 95% confidence intervals for each scale point.

Table 4.6. *Means, Standard Errors, and 95% Confidence Intervals For the Combined Reading Score, and for the Three Reading Process Subscales, For Each Point on the EJCPS: Scale 5*

EJCPS	N	%	Overall	SE	CI95L	CI95U	Retrieve	SE	CI95L	CI95U	Interpret	SE	CI95L	CI95U	Reflect	SE	CI95L	CI95U
1	59	1.6	**338.6**	10.44	317.8	359.4	**329.1**	9.86	309.5	348.7	**333.1**	9.714	313.7	352.4	**355.4**	9.82	335.9	375.0
2	208	5.8	**411.8**	5.85	400.2	423.5	**406.8**	6.63	393.6	420.0	**411.2**	7.232	396.8	425.6	**423.2**	6.98	409.3	437.1
3	504	13.9	**450.3**	4.40	441.5	459.0	**444.7**	4.70	435.4	454.1	**448.7**	4.673	439.4	458.0	**461.5**	4.56	452.5	470.6
4	299	8.3	**485.7**	4.26	477.2	494.2	**482.4**	4.53	473.4	491.4	**483.3**	4.574	474.2	492.4	**495.9**	4.03	487.8	503.9
5	63	1.7	**523.6**	8.79	506.1	541.1	**523.1**	10.75	501.7	544.5	**520.0**	9.848	500.4	539.6	**529.6**	8.76	512.1	547.0
6	1	0.0	**509.6**	15.94	477.9	541.3	**445.6**	41.80	362.4	528.8	**513.2**	27.19	459.1	567.3	**505.2**	20.18	465.1	545.4
7	40	1.1	**477.6**	11.79	454.1	501.1	**485.0**	11.94	461.2	508.8	**482.0**	11.51	459.1	504.9	**490.8**	9.94	471.0	510.6
8	575	15.9	**526.5**	4.13	518.3	534.7	**525.5**	4.26	517.0	534.0	**525.4**	4.357	516.7	534.0	**531.3**	4.75	521.8	540.7
9	1083	29.9	**561.6**	2.34	556.9	566.3	**561.1**	2.52	556.1	566.1	**563.0**	2.519	557.9	568.0	**567.0**	2.53	562.0	572.1
10	617	17.1	**602.6**	2.77	597.1	608.1	**600.5**	3.45	593.7	607.4	**603.0**	3.252	596.5	609.5	**604.3**	2.76	598.8	609.8
11	167	4.6	**642.4**	4.81	632.9	652.0	**639.1**	5.43	628.3	649.9	**644.9**	4.787	635.4	654.4	**642.8**	4.87	633.1	652.5
All available	3616	93.8	**531.1**	3.06	525.0	537.2	**528.8**	3.07	522.7	534.9	**530.9**	3.136	524.7	537.2	**537.3**	2.93	531.5	543.2
Missing	238	6.2	**459.7**	11.14	437.5	481.8	**455.8**	11.21	433.5	478.1	**459.0**	10.76	437.6	480.4	**469.7**	10.49	448.9	490.6
Total	3854	100.0	**526.7**	3.24	520.2	533.1	**524.3**	3.25	517.8	530.8	**526.5**	3.287	519.9	533.0	**533.2**	3.10	527.0	539.3

Table 4.7. *Means, Standard Errors, and 95% Confidence Intervals For the Combined Reading Score, and for the Three Reading Process Subscales, For Each Point on the EJCPS: Scale 6*

EJCPS	N	%	Overall	SE	CI95L	CI95U	Retrieve	SE	CI95L	CI95U	Interpret	SE	CI95L	CI95U	Reflect	SE	CI95L	CI95U
1	59	1.6	**338.6**	10.44	317.8	359.4	**329.1**	9.86	309.5	348.7	**333.1**	9.714	313.7	352.4	**355.4**	9.82	335.9	375.0
2	208	5.8	**411.8**	5.85	400.2	423.5	**406.8**	6.63	393.6	420.0	**411.2**	7.232	396.8	425.6	**423.2**	6.98	409.3	437.1
3	504	13.9	**450.3**	4.40	441.5	459.0	**444.7**	4.70	435.4	454.1	**448.7**	4.673	439.4	458.0	**461.5**	4.56	452.5	470.6
4	299	8.3	**485.7**	4.26	477.2	494.2	**482.4**	4.53	473.4	491.4	**483.3**	4.574	474.2	492.4	**495.9**	4.03	487.8	503.9
5	64	1.8	**523.4**	8.71	506.1	540.8	**522.1**	10.73	500.8	543.5	**519.9**	9.814	500.4	539.5	**529.3**	8.70	511.9	546.6
6	40	1.1	**477.6**	11.79	454.1	501.1	**485.0**	11.94	461.2	508.8	**482.0**	11.51	459.1	504.9	**490.8**	9.94	471.0	510.6
7	575	15.9	**526.5**	4.13	518.3	534.7	**525.5**	4.26	517.0	534.0	**525.4**	4.357	516.7	534.0	**531.3**	4.75	521.8	540.7
8	1083	29.9	**561.6**	2.34	556.9	566.3	**561.1**	2.52	556.1	566.1	**563.0**	2.519	557.9	568.0	**567.0**	2.53	562.0	572.1
9	617	17.1	**602.6**	2.77	597.1	608.1	**600.5**	3.45	593.7	607.4	**603.0**	3.252	596.5	609.5	**604.3**	2.76	598.8	609.8
10	167	4.6	**642.4**	4.81	632.9	652.0	**639.1**	5.43	628.3	649.9	**644.9**	4.787	635.4	654.4	**642.8**	4.87	633.1	652.5
All available	3616	93.8	**531.1**	3.06	525.0	537.2	**528.8**	3.07	522.7	534.9	**530.9**	3.136	524.7	537.2	**537.3**	2.93	531.5	543.2
Missing	238	6.2	**459.7**	11.14	437.5	481.8	**455.8**	11.21	433.5	478.1	**459.0**	10.76	437.6	480.4	**469.7**	10.49	448.9	490.6
Total	3854	100.0	**526.7**	3.24	520.2	533.1	**524.3**	3.25	517.8	530.8	**526.5**	3.287	519.9	533.0	**533.2**	3.10	527.0	539.3

Table 4.8. Means, Standard Errors, and 95% Confidence Intervals For the Combined Reading Score, and for the Three Reading Process Subscales, For Each Point on the EJCPS: Scale 7

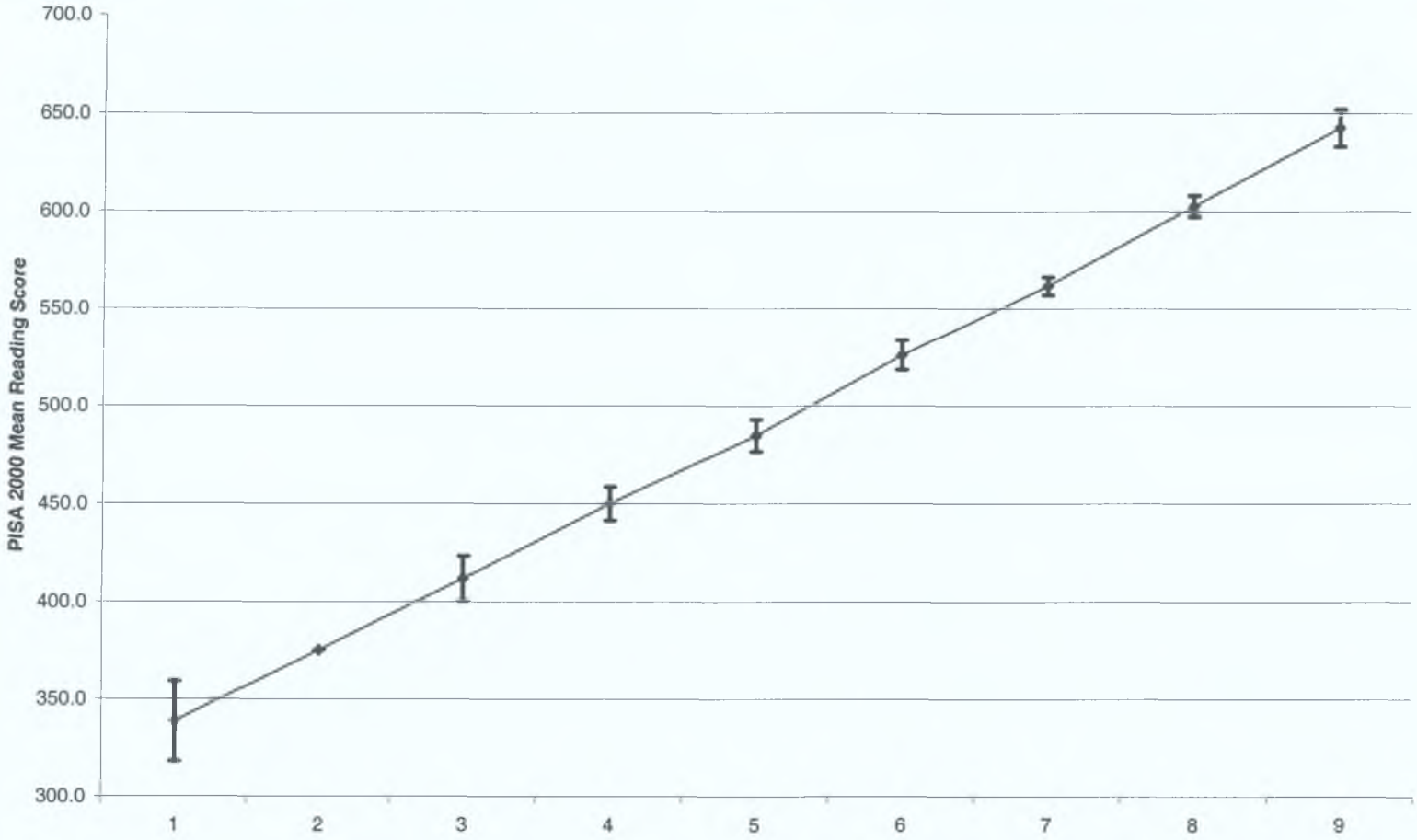
EJCPS	N	%	Overall	SE	CI95L	CI95U	Retrieve	SE	CI95L	CI95U	Interpret	SE	CI95L	CI95U	Reflect	SE	CI95L	CI95U
1	59	1.6	**338.6**	10.44	317.8	359.4	**329.1**	9.86	309.5	348.7	**333.1**	9.714	313.7	352.4	**355.4**	9.82	335.9	375.0
2	208	5.8	**411.8**	5.85	400.2	423.5	**406.8**	6.63	393.6	420.0	**411.2**	7.232	396.8	425.6	**423.2**	6.98	409.3	437.1
3	504	13.9	**450.3**	4.40	441.5	459.0	**444.7**	4.70	435.4	454.1	**448.7**	4.673	439.4	458.0	**461.5**	4.56	452.5	470.6
4	300	8.3	**485.8**	4.24	477.3	494.2	**482.3**	4.50	473.3	491.2	**483.4**	4.545	474.3	492.4	**495.9**	4.02	487.9	503.9
5	102	2.8	**505.8**	7.56	490.7	520.8	**508.4**	8.76	490.9	525.8	**505.3**	8.347	488.7	521.9	**514.5**	7.02	500.6	528.5
6	575	15.9	**526.5**	4.13	518.3	534.7	**525.5**	4.26	517.0	534.0	**525.4**	4.357	516.7	534.0	**531.3**	4.75	521.8	540.7
7	1083	29.9	**561.6**	2.34	556.9	566.3	**561.1**	2.52	556.1	566.1	**563.0**	2.519	557.9	568.0	**567.0**	2.53	562.0	572.1
8	617	17.1	**602.6**	2.77	597.1	608.1	**600.5**	3.45	593.7	607.4	**603.0**	3.252	596.5	609.5	**604.3**	2.76	598.8	609.8
9	167	4.6	**642.4**	4.81	632.9	652.0	**639.1**	5.43	628.3	649.9	**644.9**	4.787	635.4	654.4	**642.8**	4.87	633.1	652.5
All available	3616	93.8	**531.1**	3.06	525.0	537.2	**528.8**	3.07	522.7	534.9	**530.9**	3.136	524.7	537.2	**537.3**	2.93	531.5	543.2
Missing	238	6.2	**459.7**	11.14	437.5	481.8	**455.8**	11.21	433.5	478.1	**459.0**	10.76	437.6	480.4	**469.7**	10.49	448.9	490.6
Total	3854	100.0	**526.7**	3.24	520.2	533.1	**524.3**	3.25	517.8	530.8	**526.5**	3.287	519.9	533.0	**533.2**	3.10	527.0	539.3

Table 4.9. Pearson Correlations Between Eight EJCPS Scales and PISA 2000 Reading: Combined Scale and Process Subscales

	EJCPS Scale							
PISA Scale	Scale 1	Scale 2	Scale 3	Scale 4	Scale 5	Scale 6	Scale 7	Scale 8
Combined Scale	.720	.720	.717	.718	.691	.701	.712	.720
Retrieve	.690	.690	.687	.689	.665	.674	.683	.690
Interpret	.703	.703	.700	.701	.677	.686	.700	.703
Reflect	.709	.709	.706	.707	.680	.690	.700	.709

Correlations are all significant (p<.001).

Figure 4.2. Preferred 9-Point EJCPS Plotted Against PISA 2000 Mean Combined Reading Scores (and their 95% Confidence Intervals)



Note. Scale point 2 is a placeholder calculated as the midpoint between 1 and 3 and therefore has no standard error or confidence interval associated with it.

4.4.4. Analyses of the Sub-Cohort of PISA 2000 students who Took the Junior Certificate in 2000

Since the PISA 2000 cohort includes students taking the Junior Certificate in 1999 and 2000, there is a possibility that there are systematic differences between the two groups, either in terms of the characteristics of students that, in turn, relate to achievement; or in terms of the Junior Certificate English examination (e.g., the application of the marking schemes, the questions asked and the influence these had on examinee choice). To account for these, the analyses in sections 4.4.2 and 4.4.3 were replicated only for the sub-cohort taking the Junior Certificate English examination in 2000 (the same year as PISA) and the results, shown in Tables A4.1 to A4.7 in Appendix 4, confirm the choice of scale 8 as the preferred way to scale the EJCPS. The only notable difference is that the students taking the examination in 2000 score slightly lower on the PISA reading scales (by about one-sixth of a standard deviation, on average). The variation in achievement of the two groups is similar: for all students participating in PISA 2000 and for whom EJCPS data are available, the standard deviation is 90.8; for the sub-cohort taking Junior Certificate English in 2000, it is 92.1.[34] Table A4.8 shows the Pearson correlations between achievement on the PISA reading scales and the eight EJCPS scales explored for the sub-cohort. The pattern of correlations is very similar to that observed for all students participating in PISA 2000 (Table 4.10), although the correlations are marginally higher.

4.4.5. Confirmatory Analyses Using the PISA 2003 Reading / Junior Certificate English Data

To account for differences in the PISA 2000 and PISA 2003 samples that might have arisen from sampling fluctuation or differences in examinee behaviour or marking of the Junior Certificate English examination across the two PISA cycles, the mean scores associated with scale 3 (Table 4.4) (which is the same as the final scale, scale 8, with the exception of having a two-point gap between the lowest and second-lowest groups), were computed for the PISA 2003 cohort – both for all PISA 2003 students, and for the sub-cohort of PISA 2003 students attempting the Junior Certificate English examination in 2003. Means for the combined scale only were computed, since reading subscales are not available for PISA 2003. The Pearson correlations associated with these were also

[34] These standard deviations are the weighted averages of the individual standard deviations associated with the five plausible values for combined reading.

computed. Results are shown in Table A4.9. Comparing the magnitude of the correlation coefficient for scale 3 (.672 for all PISA 2003 students), it is almost identical to the correlation reported for the original 12-point EJCPS (.673; Cosgrove et al., 2005, Table 6.9). However, the intervals between EJCPS scale points differ somewhat. For example, while the PISA reading score differences between scale points 4, 5, 6, 7, 8 are comparable, the pattern is different at the lower end of the scale. While, in the 2000 dataset, there is a large gap between the lowest and second lowest scale points (about 73 points), this is notably smaller in 2003 (48 scale points). This may reflect comparatively better performance of the lowest achievers, somewhat more stringent marking on the JCE, or a mixture of these. What the comparison does demonstrate is that the EJCPS scaling appears to be more stable across time at the middle and upper regions compared to the lower region.

4.4.6. The Preferred EJCPS

Due to the dips in the mid-upper regions of scales 5 and 6 and the overlap observed in scale 7 between EJCPS groups 9 and 10, it was decided to select scale 8 as the preferred scale for the EJCPS. This is a 9-point scale with a 1-grade overlap between ordinary and foundation, a 3-grade overlap between higher and ordinary, and 2 scale points between the lowest and second lowest groups.

There are three immediate implications of the analyses presented. First, foundation-level students obtaining grades B, C, D, E and F (46 students), and ordinary-level students obtaining grades E and F (13 students), cannot be distinguished from one another on the PISA reading scales. This small group of students forms a single low-achieving mass at the lower tail of the EJCPS. Second, the analyses provide support for the 3-grade overlap between higher and ordinary levels first suggested by Martin and Hickey (1992) and used by Shiel et al. (2001) and Cosgrove et al. (2005). Third, there is no evidence of 'stretching' at the upper end of the achievement scale, which one might expect on public examinations (e.g., Greaney & Kellaghan, 1996); mean score differences are fairly even between scale points here.

4.5. Results: Junior Certificate Performance Scale for Mathematics (MJCPS)

4.5.1. Visual Exploratory Analyses

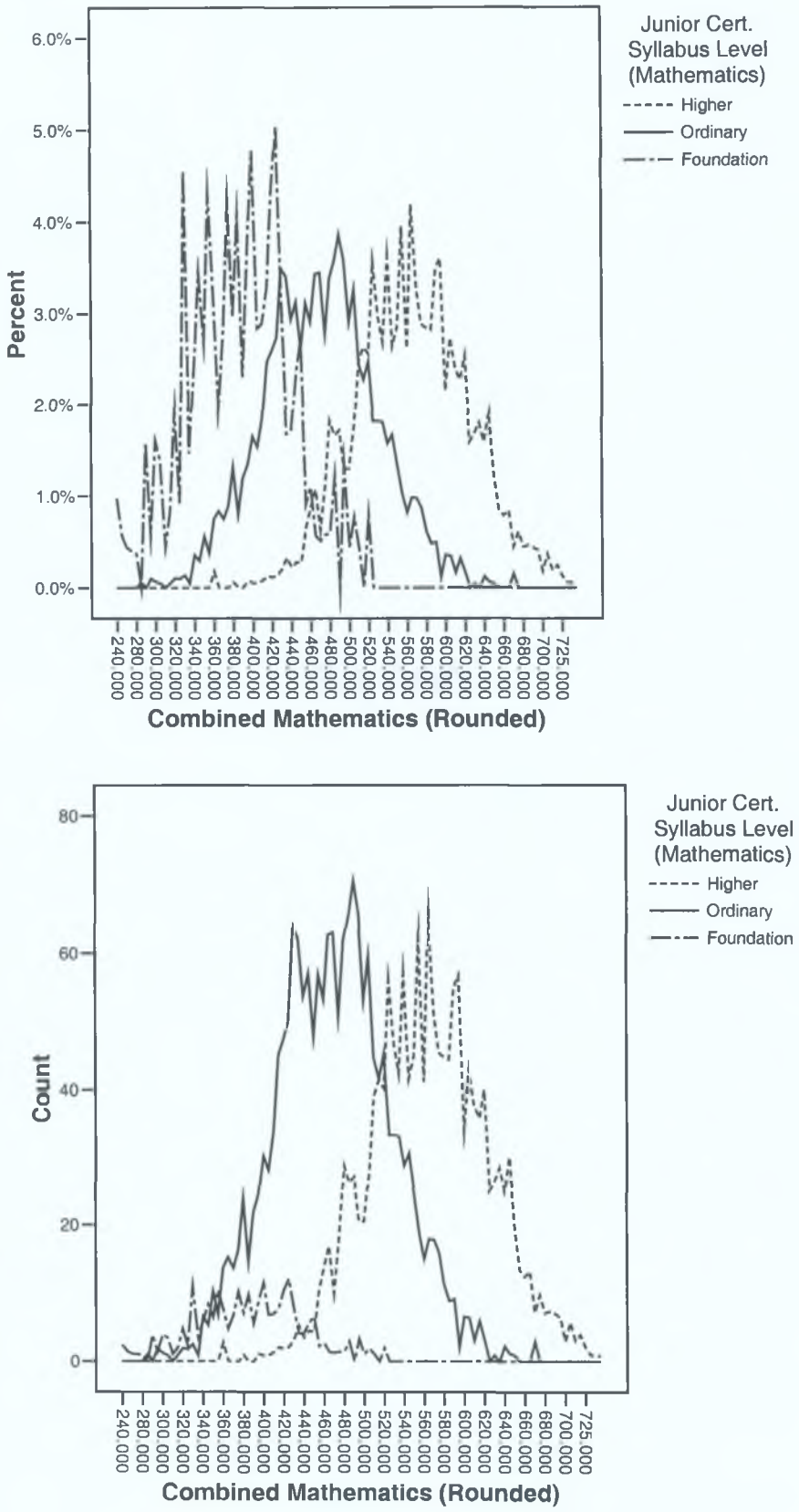
Figure 4.3 shows the frequency distribution of PISA 2003 combined mathematics scores by syllabus level, expressed in terms of both numbers of cases and in terms of the percentage of all cases within syllabus level.

The distribution of percentages at foundation level is noticeably more uneven than at higher or ordinary level. Unlike Junior Certificate English, the distributions are not notably skewed (skewness for higher, ordinary and foundation levels, respectively = .022, .084 and -.089) The mean scores of students on PISA 2003 mathematics at the three levels of the Junior Certificate mathematics examination are 563.0 (SE = 2.1), 469.1 (SE = 2.0) and 385.4 (SE = 5.2)[35], and these correspond, respectively, to the 76th, 34th, and 8th percentile points of the overall achievement distribution.

There is more of an overlap between the distributions at higher and ordinary levels than between ordinary and foundation levels (as with English/reading). The intersection between the higher and ordinary level achievement distributions is around 540 score points which corresponds to the 35th percentile of the higher-level distribution, and the 88th percentile at ordinary level. The intersection between the ordinary and foundation level achievement distributions is around 420 score points which corresponds to the 19th percentile of the ordinary level distribution and the 72nd percentile at foundation level. The achievement distributions for higher and foundation levels overlap at around 460 score points, which corresponds to the 93rd percentile at foundation level, and between the 3rd and 4th percentile at higher level.

[35] As with section 4.4.1, the percentile points and scale scores referred to in this section are approximate and are based on the mean of the five plausible values and computed in SPSS. Again, the exceptions to this are the mean score estimates for Higher, Ordinary and foundation levels, which are taken from Cosgrove et al. (2005, Table 4.24) and were computed in WesVar, taking both sampling and measurement error into account.

Figure 4.3. *Frequency Distribution of PISA 2003 Combined Mathematics Scores by Higher, Ordinary and foundation Level Mathematics (a) Percent and (b) Number (Count) of Cases Within Syllabus Level*



4.5.2. Exploration of Mean Mathematics Scores Associated With Various Versions of the MJCPS: PISA 2003 Cohort

Table 4.10 shows the eight Junior Certificate mathematics scales (MJCPS) explored using the PISA 2003 mathematics achievement data. As with the analyses for English/reading, for each of the eight scales, the combined mathematics scores, as well as scores on the four mathematics subscales are compared, using 95% confidence intervals to ascertain the extent of overlap between each point on the MJCPS.

Table 4.11 shows the mean scores of all students participating in PISA 2003 for the combined mathematics scale, and on the four mathematics subscales at each point on the original 12-point MJCPS. The table shows, first, that only 23 students are at the lowest three points on the scale, and that no student is at the lowest point.

Second, using the 95% confidence intervals as a guide, one cannot distinguish between the PISA scores of students at MJCPS scale points 2, 3 and 4. There is also an overlap in the 95% confidence intervals for all scales between scale points 4 and 5. For the remainder of scale points, there is no overlap between adjacent points.

Third, apart from the lowest four categories, there is a minimum of 217 students at each point on the scale, and the associated standard errors are smaller for these groups. The overlapping at the lower end of the scale relates to the size of the standard errors that one would expect from such small groups of students.

Fourth, the 'clustering' at the lower end of the scale is not in evidence to the same degree for mathematics as it was for English; nor is there evidence of 'clumping' of students at a particular point on the scale as was observed with scale point 9 on the original EJCPS. Finally, the mean scores of students at each point on the MJCPS on each subscale is consistent with overall mean performance (surprising, perhaps, given that the material in Chapter 2 suggests that students were expected to have somewhat differing levels of familiarity with the mathematics subscales; however it was also pointed out that a cross-sectional design is not the optimal way to assess the association between curricular content and achievement).

Table 4.10. Description of MJCPS Scoring Schemes Explored

Scale 1	Grade/Level			Scale 5	Grade/Level		
Scale point	Higher	Ordinary	Foundation	Scale point	Higher	Ordinary	Foundation
12	A						
11	B			11	A		
10	C			10	B		
9	D	A		9	C		
8	E	B		8	D		
7	F	C		7	E		
6		D	A	6	F		
5		E	B	5		A	
4		F	C	4		B	
3			D	3		C	
2			E	2		D	A
1			F	1		E, F	B, C, D, E, F

Scale 2	Grade/Level			Scale 6	Grade/Level		
Scale point	Higher	Ordinary	Foundation	Scale point	Higher	Ordinary	Foundation
				10	A		
9	A			9	B		
8	B			8	C		
7	C			7	D		
6	D	A		6	E		
5	E	B		5	F	A	
4	F	C		4		B	
3		D	A	3		C	
2		E	B	2		D	A
1		F	C, D, E, F	1		E, F	B, C, D, E, F

Scale 3	Grade/Level			Scale 7	Grade/Level		
Scale point	Higher	Ordinary	Foundation	Scale point	Higher	Ordinary	Foundation
				9	A		
8	A			8	B		
7	B			7	C		
6	C			6	D		
5	D	A		5	E	A	
4	E	B		4	F	B	
3	F	C		3		C	
2		D	A	2		D	A
1		E, F	B, C, D, E, F	1		E, F	B, C, D, E, F

Scale 4	Grade/Level			Scale 8 - Final Scale	Grade/Level		
Scale point	Higher	Ordinary	Foundation	Scale point	Higher	Ordinary	Foundation
8	A			10	A		
7	B			8	B		
6	C			7	C		
5	D	A		6	D	A	
4	E	B		5	E	B	
3	F	C		4	F	C	
2		D, E	A, B	3		D	A
1		F	C, D, E, F	1		E, F	B, C, D, E, F

Description of scales

1 Original 12-point scale

2 9-point scale with original points 4, 3, 2, 1 collapsed into one category

3 8-point scale with original points 5, 4, 3, 2, 1 collapsed into one category

4 8-point scale with original points 4, 3, 2, 1 collapsed into one category and original 5 recoded to 6

5 11-point scale with no overlap between Higher and Ordinary level grades

6 10-point scale with a one-point overlap between Higher and Ordinary level grades

7 9-point scale with a two-point overlap between Higher and Ordinary level grades

8 8-point scale with original points 5, 4, 3, 2, 1 collapsed into one category and two scale points separating the lowest and second lowest categories and the highest and second highest categories

Table 4.11. Means, Standard Errors, and 95% Confidence Intervals For the Combined Mathematics Score, and for the Four Mathematics Content Area Subscales, For Each Point on the MCJPS: Scale 1

MJCPS	N	%	Overall	SE	CI95L	CI95U	S&S	SE	CI95L	CI95U	C&R	SE	CI95L	CI95U	U	SE	CI95L	CI95U	Q	SE	CI95L	CI95U
1*	0	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
2	6	0.1	**300.5**	29.44	242.8	358.2	**271.9**	34.72	203.8	339.9	**274.6**	31.60	212.7	336.5	**318.7**	27.50	264.8	372.6	**292.9**	37.10	220.2	365.6
3	17	0.4	**352.6**	15.13	322.9	382.2	**315.2**	18.54	278.8	351.5	**347.4**	17.11	313.9	381.0	**358.2**	16.76	325.3	391.1	**350.1**	15.30	320.1	380.1
4	81	2.1	**384.2**	6.80	370.9	397.5	**362.5**	8.75	345.4	379.7	**384.3**	8.55	367.5	401.0	**402.8**	8.30	386.6	419.1	**382.5**	8.57	365.7	399.3
5	217	5.6	**401.3**	4.45	392.6	410.1	**374.8**	5.88	363.3	386.4	**400.3**	5.59	389.3	411.2	**414.7**	4.80	405.3	424.1	**395.9**	4.96	386.2	405.6
6	419	10.8	**433.4**	4.11	425.4	441.5	**407.6**	4.71	398.4	416.8	**437.1**	4.40	428.4	445.7	**444.0**	4.43	435.3	452.6	**433.0**	4.95	423.3	442.7
7	604	15.6	**466.6**	2.77	461.2	472.0	**439.7**	3.48	432.9	446.5	**471.8**	3.04	465.9	477.8	**479.4**	3.29	472.9	485.9	**467.0**	3.11	460.9	473.1
8	631	16.3	**497.5**	2.64	492.3	502.7	**470.4**	3.44	463.6	477.1	**500.0**	2.90	494.3	505.7	**511.4**	2.89	505.7	517.1	**495.3**	3.40	488.6	501.9
9	470	12.1	**528.7**	3.08	522.6	534.7	**500.3**	3.63	493.2	507.4	**533.0**	3.13	526.9	539.2	**544.2**	3.46	537.4	551.0	**527.1**	3.21	520.8	533.4
10	459	11.8	**551.2**	2.84	545.6	556.8	**524.4**	3.87	516.8	531.9	**554.9**	2.96	549.1	560.7	**567.8**	2.96	562.0	573.6	**550.7**	3.11	544.6	556.8
11	490	12.6	**577.4**	3.06	571.4	583.4	**550.2**	3.81	542.7	557.6	**581.7**	2.57	576.6	586.7	**593.8**	2.65	588.6	599.0	**578.0**	3.32	571.5	584.5
12	249	6.4	**623.2**	4.46	614.4	631.9	**601.1**	6.21	589.0	613.3	**623.5**	3.80	616.0	630.9	**641.0**	4.43	632.3	649.7	**619.8**	4.04	611.9	627.7
All available	3642	93.9	**505.9**	2.45	501.1	510.7	**476.2**	2.43	471.5	481.0	**509.1**	2.44	504.3	513.9	**520.4**	2.63	515.3	525.6	**501.7**	2.48	496.8	506.5
Missing	238	6.1	**456.3**	9.48	437.7	474.9	**430.1**	9.79	411.0	449.3	**457.8**	10.39	437.5	478.2	**467.9**	10.42	447.5	488.3	**455.1**	9.16	437.1	473.0
Total	3880	100.0	**502.8**	2.45	498.0	507.6	**476.2**	2.43	471.5	481.0	**506.0**	2.45	501.2	510.7	**517.2**	2.65	512.0	522.4	**504.7**	2.48	499.9	509.6

*In the PISA 2003 cohort, no student was awarded Grade F at Foundation Level

Note. 'S&S' = Space & Shape subscale; 'C&R' = Change & Relationships subscale; 'U' = Uncertainty subscale; 'Q' = Quantity subscale.

Table 4.12 shows mean PISA mathematics scores for the second MJCPS – i.e., a 9-point scale that results from collapsing EJCPS score categories 2, 3, and 4. The scale is now smoothed somewhat, with a PISA combined mathematics score difference of between about 22 and 32 scale points at each MJCPS point. An exception to this is the mean score difference between students at the second highest and highest MJCPS scale points (e.g., about 46 PISA scale points for the combined mathematics scale). An examination of the 95% confidence intervals suggests that the mean score differences at the second lowest and lowest points are significantly different only in the case of the combined mathematics scale and the Change & Relationships subscale; for the other three subscales, the confidence intervals for these two groups overlap. Scales 3 and 4 thus explore two alternative methods of collapsing the second-lowest group (as with English/reading): collapsing the second lowest group with the lowest (scale 3) and collapsing the second-lowest group with the third-lowest (scale 4).

Table 4.13 shows the mean scores for the combined mathematics scale and four subscales for scale 3. The 8-point scale has empirically distinct mean scores for both the combined scale and each of the four subscales at each point (using the 95% confidence intervals as a guide). However, the mean score differences between the lowest and second lowest groups (ranging between about 37 to 46 scale points depending on the scale considered), and between the second highest and highest groups (ranging between about 42 to 51 scale points) are somewhat larger than mean score differences between other adjacent groups, and larger also for the mean score differences associated with scale 2.

Table 4.14 shows the mean scores for scale 4. Collapsing the second lowest MJCPS group upwards rather than downwards results in a somewhat larger gap between the new lowest and second lowest groups of about 42 to 52 scale points for the combined mathematics scale (for example). It was decided, therefore, to proceed with scale 3 and explore various ways of overlapping the MJCPS scale at higher and ordinary level.

Table 4.12. Means, Standard Errors, and 95% Confidence Intervals For the Combined Mathematics Score, and for the Four Mathematics Content Area Subscales, For Each Point on the MCJPS: Scale 2

MJCPS	N	%	Overall	SE	CI95L	CI95U	S&S	SE	CI95L	CI95U	C&R	SE	CI95L	CI95U	U	SE	CI95L	CI95U	Q	SE	CI95L	CI95U
1	104	2.7	**374.6**	6.91	361.1	388.2	**350.1**	7.90	334.6	365.5	**372.5**	8.46	355.9	389.1	**391.1**	7.84	375.8	406.5	**372.5**	8.00	356.8	388.2
2	217	5.6	**401.3**	4.45	392.6	410.1	**374.8**	5.88	363.3	386.4	**400.3**	5.59	389.3	411.2	**414.7**	4.80	405.3	424.1	**395.9**	4.96	386.2	405.6
3	419	10.8	**433.4**	4.11	425.4	441.5	**407.6**	4.71	398.4	416.8	**437.1**	4.40	428.4	445.7	**444.0**	4.43	435.3	452.6	**433.0**	4.95	423.3	442.7
4	604	15.6	**466.6**	2.77	461.2	472.0	**439.7**	3.48	432.9	446.5	**471.8**	3.04	465.9	477.8	**479.4**	3.29	472.9	485.9	**467.0**	3.11	460.9	473.1
5	631	16.3	**497.5**	2.64	492.3	502.7	**470.4**	3.44	463.6	477.1	**500.0**	2.90	494.3	505.7	**511.4**	2.89	505.7	517.1	**495.3**	3.40	488.6	501.9
6	470	12.1	**528.7**	3.08	522.6	534.7	**500.3**	3.63	493.2	507.4	**533.0**	3.13	526.9	539.2	**544.2**	3.46	537.4	551.0	**527.1**	3.21	520.8	533.4
7	459	11.8	**551.2**	2.84	545.6	556.8	**524.4**	3.87	516.8	531.9	**554.9**	2.96	549.1	560.7	**567.8**	2.96	562.0	573.6	**550.7**	3.11	544.6	556.8
8	490	12.6	**577.4**	3.06	571.4	583.4	**550.2**	3.81	542.7	557.6	**581.7**	2.57	576.6	586.7	**593.8**	2.65	588.6	599.0	**578.0**	3.32	571.5	584.5
9	249	6.4	**623.2**	4.46	614.4	631.9	**601.1**	6.21	589.0	613.3	**623.5**	3.80	616.0	630.9	**641.0**	4.43	632.3	649.7	**619.8**	4.04	611.9	627.7
All available	3642	93.9	**505.9**	2.45	501.1	510.7	**476.2**	2.43	471.5	481.0	**509.1**	2.44	504.3	513.9	**520.4**	2.63	515.3	525.6	**501.7**	2.48	496.8	506.5
Missing	238	6.1	**456.3**	9.48	437.7	474.9	**430.1**	9.79	411.0	449.3	**457.8**	10.39	437.5	478.2	**467.9**	10.42	447.5	488.3	**455.1**	9.16	437.1	473.0
Total	3880	100.0	**502.8**	2.45	498.0	507.6	**476.2**	2.43	471.5	481.0	**506.0**	2.45	501.2	510.7	**517.2**	2.65	512.0	522.4	**504.7**	2.48	499.9	509.6

Note. 'S&S' = Space & Shape subscale; 'C&R' = Change & Relationships subscale; 'U' = Uncertainty subscale; 'Q' = Quantity subscale.

Table 4.13. Means, Standard Errors, and 95% Confidence Intervals For the Combined Mathematics Score, and for the Four Mathematics Content Area Subscales, For Each Point on the MCJPS: Scale 3

MJCPS	N	%	Overall	SE	CI95L	CI95U	S&S	SE	CI95L	CI95U	C&R	SE	CI95L	CI95U	U	SE	CI95L	CI95U	Q	SE	CI95L	CI95U
1	321	8.8	**392.7**	4.13	384.6	400.8	**366.8**	5.17	356.7	377.0	**391.3**	5.283	380.9	401.6	**407.1**	4.69	397.9	416.3	**388.4**	4.71	379.1	397.6
2	419	11.5	**433.4**	4.11	425.4	441.5	**407.6**	4.71	398.4	416.8	**437.1**	4.4	428.4	445.7	**444.0**	4.43	435.3	452.6	**433.0**	4.95	423.3	442.7
3	604	16.6	**466.6**	2.77	461.2	472.0	**439.7**	3.48	432.9	446.5	**471.8**	3.044	465.9	477.8	**479.4**	3.29	472.9	485.9	**467.0**	3.11	460.9	473.1
4	631	17.3	**497.5**	2.64	492.3	502.7	**470.4**	3.44	463.6	477.1	**500.0**	2.9	494.3	505.7	**511.4**	2.89	505.7	517.1	**495.3**	3.40	488.6	501.9
5	470	12.9	**528.7**	3.08	522.6	534.7	**500.3**	3.63	493.2	507.4	**533.0**	3.128	526.9	539.2	**544.2**	3.46	537.4	551.0	**527.1**	3.21	520.8	533.4
6	459	12.6	**551.2**	2.84	545.6	556.8	**524.4**	3.87	516.8	531.9	**554.9**	2.955	549.1	560.7	**567.8**	2.96	562.0	573.6	**550.7**	3.11	544.6	556.8
7	490	13.5	**577.4**	3.06	571.4	583.4	**550.2**	3.81	542.7	557.6	**581.7**	2.57	576.6	586.7	**593.8**	2.65	588.6	599.0	**578.0**	3.32	571.5	584.5
8	249	6.8	**623.2**	4.46	614.4	631.9	**601.1**	6.21	589.0	613.3	**623.5**	3.804	616.0	630.9	**641.0**	4.43	632.3	649.7	**619.8**	4.04	611.9	627.7
All available	3642	93.9	**505.9**	2.45	501.1	510.7	**476.2**	2.43	471.5	481.0	**509.1**	2.436	504.3	513.9	**520.4**	2.63	515.3	525.6	**501.7**	2.48	496.8	506.5
Missing	238	6.1	**456.3**	9.48	437.7	474.9	**430.1**	9.79	411.0	449.3	**457.8**	10.39	437.5	478.2	**467.9**	10.42	447.5	488.3	**455.1**	9.16	437.1	473.0
Total	3880	100.0	**502.8**	2.45	498.0	507.6	**476.2**	2.43	471.5	481.0	**506.0**	2.447	501.2	510.7	**517.2**	2.65	512.0	522.4	**504.7**	2.48	499.9	509.6

Note. 'S&S' = Space & Shape subscale; 'C&R' = Change & Relationships subscale; 'U' = Uncertainty subscale; 'Q' = Quantity subscale.

Table 4.14. Means, Standard Errors, and 95% Confidence Intervals For the Combined Mathematics Score, and for the Four Mathematics Content Area Subscales, For Each Point on the MCJPS: Scale 4

MJCPS	*N*	*%*	*Overall*	*SE*	*CI95L*	*CI95U*	*S&S*	*SE*	*CI95L*	*CI95U*	*C&R*	*SE*	*CI95L*	*CI95U*	*U*	*SE*	*CI95L*	*CI95U*	*Q*	*SE*	*CI95L*	*CI95U*
1	104	2.9	**374.6**	6.91	361.1	388.2	**350.1**	7.90	334.6	365.5	**372.5**	8.46	355.9	389.1	**391.1**	7.84	375.8	406.5	**372.5**	8.00	356.8	388.2
2	636	17.5	**422.5**	3.00	416.6	428.4	**396.4**	3.56	389.4	403.4	**424.5**	3.21	418.2	430.8	**434.0**	3.20	427.7	440.3	**420.4**	3.70	413.1	427.6
3	604	16.6	**466.6**	2.77	461.2	472.0	**439.7**	3.48	432.9	446.5	**471.8**	3.04	465.9	477.8	**479.4**	3.29	472.9	485.9	**467.0**	3.11	460.9	473.1
4	631	17.3	**497.5**	2.64	492.3	502.7	**470.4**	3.44	463.6	477.1	**500.0**	2.90	494.3	505.7	**511.4**	2.89	505.7	517.1	**495.3**	3.40	488.6	501.9
5	470	12.9	**528.7**	3.08	522.6	534.7	**500.3**	3.63	493.2	507.4	**533.0**	3.13	526.9	539.2	**544.2**	3.46	537.4	551.0	**527.1**	3.21	520.8	533.4
6	459	12.6	**551.2**	2.84	545.6	556.8	**524.4**	3.87	516.8	531.9	**554.9**	2.96	549.1	560.7	**567.8**	2.96	562.0	573.6	**550.7**	3.11	544.6	556.8
7	490	13.5	**577.4**	3.06	571.4	583.4	**550.2**	3.81	542.7	557.6	**581.7**	2.57	576.6	586.7	**593.8**	2.65	588.6	599.0	**578.0**	3.32	571.5	584.5
8	249	6.8	**623.2**	4.46	614.4	631.9	**601.1**	6.21	589.0	613.3	**623.5**	3.80	616.0	630.9	**641.0**	4.43	632.3	649.7	**619.8**	4.04	611.9	627.7
All available	3642	93.9	**505.9**	2.45	501.1	510.7	**476.2**	2.43	471.5	481.0	**509.1**	2.44	504.3	513.9	**520.4**	2.63	515.3	525.6	**501.7**	2.48	496.8	506.5
Missing	238	6.1	**456.3**	9.48	437.7	474.9	**430.1**	9.79	411.0	449.3	**457.8**	10.39	437.5	478.2	**467.9**	10.42	447.5	488.3	**455.1**	9.16	437.1	473.0
Total	3880	100.0	**502.8**	2.45	498.0	507.6	**476.2**	2.43	471.5	481.0	**506.0**	2.45	501.2	510.7	**517.2**	2.65	512.0	522.4	**504.7**	2.48	499.9	509.6

Note. 'S&S' = Space & Shape subscale; 'C&R' = Change & Relationships subscale; 'U' = Uncertainty subscale; 'Q' = Quantity subscale.

Scales 5, 6 and 7 explore the consequences of having no overlap, a one-grade overlap, and a two-grade overlap, respectively, between higher and ordinary levels. The mean PISA mathematics scores associated with each of these is shown in Tables 4.15, 4.16 and 4.17. Similar to the patterns observed for English/reading, scales 5 and 6 suggest a dip in achievement on PISA around the middle of the MJCPS (i.e., students obtaining grades E and F at higher level score lower than students obtaining grade A at ordinary level). A comparison of scale 7 and scale 3, which are identical except that scale 7 has a 2-grade rather than a 3-grade overlap between higher and ordinary, suggests that the original 3-grade overlap is superior, given that the 95% confidence intervals for PISA mean scores associated with MJCPS score 5 and 6 overlap.

As noted previously, the intervals between the lowest and second lowest points on the scale, and the second highest and highest points, are larger than the intervals between the other points on the scale. Therefore, scale 3 was re-scaled to range from 1 to 10, with a two-point gap between the lowest and second lowest points of the scale, and a two-point gap between the second highest and highest points on the scale. Thus the average gap between scale points on the preferred scale (scale 8) is about 29 PISA scale points, or one-third of a (national) standard deviation.

4.5.3. Exploration of Pearson Correlations Associated With Various Versions of the MJCPS: PISA 2003 Cohort

Table 4.18 shows the Pearson correlations between the eight MJCPS scales explored and achievement on PISA 2003 combined mathematics and on the four mathematics subscales. As with English/reading, there is little or no difference in the strength of correlation obtained, regardless of the MJCPS scale or the PISA scale considered. All correlations range between .66 and .75. There is a marginal decrease in the strength of the correlation as the overlap between higher and ordinary levels is decreased. The pattern of correlations supports the choice of scale 8, although the strength of its correlations with achievement is identical to two decimal places to several of the other scales explored. The advantage of scale 8 however, as noted, is that it is smoother; also, the 95% confidence intervals between adjacent groups do not overlap (as with scale 8 for EJCPS). Figure 4.4 depicts MJCPS scale 8 graphically, with the PISA combined mean mathematics score and 95% confidence intervals for each scale point.

Table 4.15. Means, Standard Errors, and 95% Confidence Intervals For the Combined Mathematics Score, and for the Four Mathematics Content Area Subscales, For Each Point on the MCJPS: Scale 5

MJCPS	N	%	Overall	SE	CI95L	CI95U	S&S	SE	CI95L	CI95U	C&R	SE	CI95L	CI95U	U	SE	CI95L	CI95U	Q	SE	CI95L	CI95U
1	321	8.8	**392.7**	4.13	384.6	400.8	**366.8**	5.17	356.7	377.0	**391.3**	5.28	380.9	401.6	**407.1**	4.69	397.9	416.3	**388.4**	4.71	379.1	397.6
2	419	11.5	**433.4**	4.11	425.4	441.5	**407.6**	4.71	398.4	416.8	**437.1**	4.40	428.4	445.7	**444.0**	4.43	435.3	452.6	**433.0**	4.95	423.3	442.7
3	589	16.2	**465.9**	2.72	460.5	471.2	**439.1**	3.58	432.1	446.1	**471.1**	3.03	465.1	477.0	**478.5**	3.30	472.0	485.0	**466.2**	3.10	460.1	472.3
4	566	15.6	**495.7**	2.67	490.5	501.0	**468.9**	3.63	461.8	476.0	**498.1**	3.02	492.1	504.0	**509.8**	3.15	503.6	516.0	**493.5**	3.76	486.1	500.8
5	170	4.7	**526.0**	4.26	517.7	534.4	**496.1**	5.29	485.8	506.5	**528.7**	4.60	519.7	537.7	**538.7**	4.70	529.5	547.9	**519.9**	4.46	511.1	528.6
6	14	0.4	**497.1**	14.41	468.9	525.4	**463.5**	15.71	432.7	494.3	**503.7**	12.98	478.2	529.1	**516.9**	15.75	486.0	547.8	**499.4**	15.54	468.9	529.8
7	65	1.8	**512.9**	9.17	494.9	530.8	**482.9**	10.21	462.9	502.9	**517.0**	10.11	497.1	536.8	**525.6**	11.04	504.0	547.3	**510.9**	9.09	493.1	528.7
8	299	8.2	**530.2**	4.07	522.2	538.2	**502.7**	4.71	493.4	511.9	**535.5**	3.77	528.1	542.9	**547.4**	4.41	538.7	556.0	**531.3**	4.18	523.1	539.5
9	459	12.6	**551.2**	2.84	545.6	556.8	**524.4**	3.87	516.8	531.9	**554.9**	2.96	549.1	560.7	**567.8**	2.96	562.0	573.6	**550.7**	3.11	544.6	556.8
10	490	13.5	**577.4**	3.06	571.4	583.4	**550.2**	3.81	542.7	557.6	**581.7**	2.57	576.6	586.7	**593.8**	2.65	588.6	599.0	**578.0**	3.32	571.5	584.5
11	249	6.8	**623.2**	4.46	614.4	631.9	**601.1**	6.21	589.0	613.3	**623.5**	3.80	616.0	630.9	**641.0**	4.43	632.3	649.7	**619.8**	4.04	611.9	627.7
All available	3642	93.9	**505.9**	2.45	501.1	510.7	**476.2**	2.43	471.5	481.0	**509.1**	2.44	504.3	513.9	**520.4**	2.63	515.3	525.6	**501.7**	2.48	496.8	506.5
Missing	238	6.1	**456.3**	9.48	437.7	474.9	**430.1**	9.79	411.0	449.3	**457.8**	10.39	437.5	478.2	**467.9**	10.42	447.5	488.3	**455.1**	9.16	437.1	473.0
Total	3880	100.0	**502.8**	2.45	498.0	507.6	**476.2**	2.43	471.5	481.0	**506.0**	2.45	501.2	510.7	**517.2**	2.65	512.0	522.4	**504.7**	2.48	499.9	509.6

Note. 'S&S' = Space & Shape subscale; 'C&R' = Change & Relationships subscale; 'U' = Uncertainty subscale; 'Q' = Quantity subscale.

Table 4.16. Means, Standard Errors, and 95% Confidence Intervals For the Combined Mathematics Score, and for the Four Mathematics Content Area Subscales, For Each Point on the MCJPS: Scale 6

MJCPS	N	%	Overall	SE	CI95L	CI95U	S&S	SE	CI95L	CI95U	C&R	SE	CI95L	CI95U	U	SE	CI95L	CI95U	Q	SE	CI95L	CI95U
1	321	8.3	**392.7**	4.13	384.6	400.8	**366.8**	5.17	356.7	377.0	**391.3**	5.28	380.9	401.6	**407.1**	4.69	397.9	416.3	**388.4**	4.71	379.1	397.6
2	419	10.8	**433.4**	4.11	425.4	441.5	**407.6**	4.71	398.4	416.8	**437.1**	4.40	428.4	445.7	**444.0**	4.43	435.3	452.6	**433.0**	4.95	423.3	442.7
3	589	15.2	**465.9**	2.72	460.5	471.2	**439.1**	3.58	432.1	446.1	**471.1**	3.03	465.1	477.0	**478.5**	3.30	472.0	485.0	**466.2**	3.10	460.1	472.3
4	566	14.6	**495.7**	2.67	490.5	501.0	**468.9**	3.63	461.8	476.0	**498.1**	3.02	492.1	504.0	**509.8**	3.15	503.6	516.0	**493.5**	3.76	486.1	500.8
5	185	4.8	**523.8**	3.99	516.0	531.6	**493.6**	5.03	483.8	503.5	**526.7**	4.39	518.1	535.3	**537.0**	4.23	528.7	545.3	**518.3**	4.31	509.8	526.7
6	65	1.7	**512.9**	9.17	494.9	530.8	**482.9**	10.21	462.9	502.9	**517.0**	10.11	497.1	536.8	**525.6**	11.04	504.0	547.3	**510.9**	9.09	493.1	528.7
7	299	7.7	**530.2**	4.07	522.2	538.2	**502.7**	4.71	493.4	511.9	**535.5**	3.77	528.1	542.9	**547.4**	4.41	538.7	556.0	**531.3**	4.18	523.1	539.5
8	459	11.8	**551.2**	2.84	545.6	556.8	**524.4**	3.87	516.8	531.9	**554.9**	2.96	549.1	560.7	**567.8**	2.96	562.0	573.6	**550.7**	3.11	544.6	556.8
9	490	12.6	**577.4**	3.06	571.4	583.4	**550.2**	3.81	542.7	557.6	**581.7**	2.57	576.6	586.7	**593.8**	2.65	588.6	599.0	**578.0**	3.32	571.5	584.5
10	249	6.4	**623.2**	4.46	614.4	631.9	**601.1**	6.21	589.0	613.3	**623.5**	3.80	616.0	630.9	**641.0**	4.43	632.3	649.7	**619.8**	4.04	611.9	627.7
All available	3642	93.9	**505.9**	2.45	501.1	510.7	**476.2**	2.43	471.5	481.0	**509.1**	2.44	504.3	513.9	**520.4**	2.63	515.3	525.6	**501.7**	2.48	496.8	506.5
Missing	238	6.1	**456.3**	9.48	437.7	474.9	**430.1**	9.79	411.0	449.3	**457.8**	10.39	437.5	478.2	**467.9**	10.42	447.5	488.3	**455.1**	9.16	437.1	473.0
Total	3880	100.0	**502.8**	2.45	498.0	507.6	**476.2**	2.43	471.5	481.0	**506.0**	2.45	501.2	510.7	**517.2**	2.65	512.0	522.4	**504.7**	2.48	499.9	509.6

Note. 'S&S' = Space & Shape subscale; 'C&R' = Change & Relationships subscale; 'U' = Uncertainty subscale; 'Q' = Quantity subscale.

Table 4.17. Means, Standard Errors, and 95% Confidence Intervals For the Combined Mathematics Score, and for the Four Mathematics Content Area Subscales, For Each Point on the MCJPS: Scale 7

MJCPS	*N*	*%*	*Overall*	*SE*	*CI95L*	*CI95U*	*S&S*	*SE*	*CI95L*	*CI95U*	*C&R*	*SE*	*CI95L*	*CI95U*	*U*	*SE*	*CI95L*	*CI95U*	*Q*	*SE*	*CI95L*	*CI95U*
1	321	8.8	**392.7**	4.13	384.6	400.8	**366.8**	5.17	356.7	377.0	**391.3**	5.28	380.9	401.6	**407.1**	4.69	397.9	416.3	**388.4**	4.71	379.1	397.6
2	419	11.5	**433.4**	4.11	425.4	441.5	**407.6**	4.71	398.4	416.8	**437.1**	4.40	428.4	445.7	**444.0**	4.43	435.3	452.6	**433.0**	4.95	423.3	442.7
3	589	16.2	**465.9**	2.72	460.5	471.2	**439.1**	3.58	432.1	446.1	**471.1**	3.03	465.1	477.0	**478.5**	3.30	472.0	485.0	**466.2**	3.10	460.1	472.3
4	581	15.9	**495.8**	2.65	490.6	501.0	**468.8**	3.54	461.8	475.7	**498.2**	2.97	492.4	504.0	**510.0**	3.10	503.9	516.0	**493.6**	3.73	486.3	500.9
5	235	6.5	**522.4**	4.19	514.2	530.6	**492.5**	5.11	482.5	502.5	**525.4**	4.56	516.5	534.4	**535.1**	4.97	525.3	544.8	**517.4**	4.24	509.1	525.7
6	299	8.2	**530.2**	4.07	522.2	538.2	**502.7**	4.71	493.4	511.9	**535.5**	3.77	528.1	542.9	**547.4**	4.41	538.7	556.0	**531.3**	4.18	523.1	539.5
7	459	12.6	**551.2**	2.84	545.6	556.8	**524.4**	3.87	516.8	531.9	**554.9**	2.96	549.1	560.7	**567.8**	2.96	562.0	573.6	**550.7**	3.11	544.6	556.8
8	490	13.5	**577.4**	3.06	571.4	583.4	**550.2**	3.81	542.7	557.6	**581.7**	2.57	576.6	586.7	**593.8**	2.65	588.6	599.0	**578.0**	3.32	571.5	584.5
9	249	6.8	**623.2**	4.46	614.4	631.9	**601.1**	6.21	589.0	613.3	**623.5**	3.80	616.0	630.9	**641.0**	4.43	632.3	649.7	**619.8**	4.04	611.9	627.7
All available	3642	93.9	**505.9**	2.45	501.1	510.7	**476.2**	2.43	471.5	481.0	**509.1**	2.44	504.3	513.9	**520.4**	2.63	515.3	525.6	**501.7**	2.48	496.8	506.5
Missing	238	6.1	**456.3**	9.48	437.7	474.9	**430.1**	9.79	411.0	449.3	**457.8**	10.39	437.5	478.2	**467.9**	10.42	447.5	488.3	**455.1**	9.16	437.1	473.0
Total	3880	100.0	**502.8**	2.45	498.0	507.6	**476.2**	2.43	471.5	481.0	**506.0**	2.45	501.2	510.7	**517.2**	2.65	512.0	522.4	**504.7**	2.48	499.9	509.6

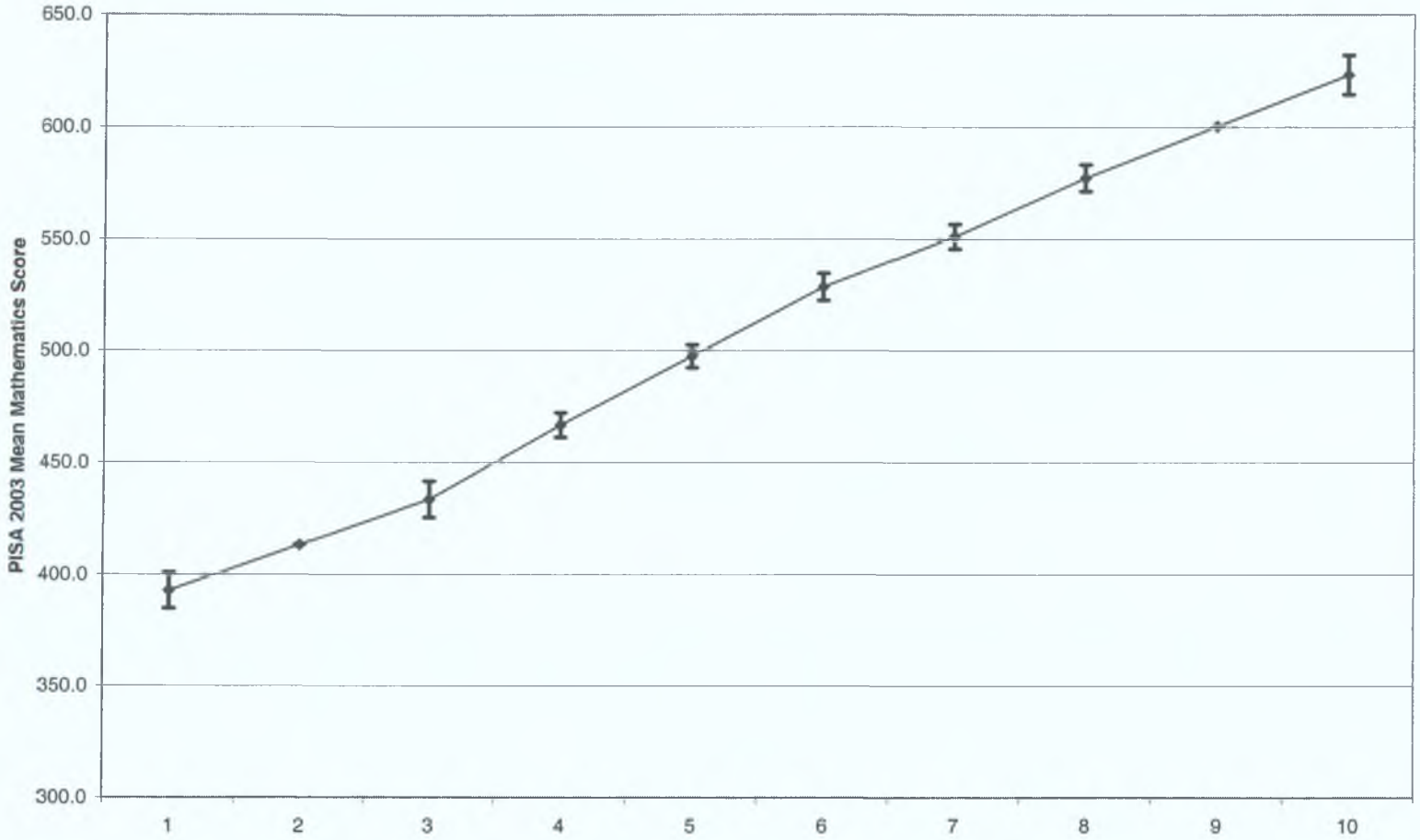
Note. 'S&S' = Space & Shape subscale; 'C&R' = Change & Relationships subscale; 'U' = Uncertainty subscale; 'Q' = Quantity subscale.

Table 4.18. Pearson Correlations Between Eight MJCPS Scales and PISA 2003 Mathematics: Combined Scale and Subscales

	MJCPS Scale							
PISA Scale	*Scale 1*	*Scale 2*	*Scale 3*	*Scale 4*	*Scale 5*	*Scale 6*	*Scale 7*	*Scale 8*
Combined Scale	.754	.753	.750	.746	.725	.735	.745	.754
Space and Shape	.681	.679	.678	.674	.655	.664	.673	.681
Change and Relationships	.741	.739	.737	.731	.713	.722	.732	.741
Uncertainty	.743	.742	.741	.737	.719	.729	.737	.743
Quantity	.731	.730	.728	.722	.706	.715	.724	.731

Correlations are all significant (p<.001).

Figure 4.4. Preferred 10-Point MJCPS Plotted Against PISA 2003 Mean Combined Mathematics Scores (and their 95% Confidence Intervals)



Note. Scale points 2 and 9 are placeholders calculated as the midpoint between 1 and 3, and 8 and 10, respectively, and therefore have no standard error or confidence interval associated with them.

4.5.4. Analyses of the Sub-Cohort of PISA 2003 students who Took the Junior Certificate in 2003

Since the PISA 2003 cohort includes students taking the Junior Certificate in 2002 and 2003, differences between the two groups, either in terms of the characteristics of students, or the Junior Certificate mathematics examination, may impact on the choice of a preferred MJCPS. As with the analyses of English/reading, to account for these, the analyses in sections 4.4.2 and 4.4.3 were replicated only for the sub-cohort taking the Junior Certificate mathematics examination in 2003 (the same year as PISA) and the results, shown in Tables A4.10 to A4.17 in Appendix 4, confirm the choice of scale 8 as the preferred scale for MJCPS. The only notable difference is that the students taking the examination in 2003 score slightly lower on the PISA mathematics scales (by about one-eighth of a standard deviation, on average). The variation in achievement of the two groups is similar (as with English/reading): for all students participating in PISA 2003 and for whom MJCPS data are available, the standard deviation is 83.50; for the sub-cohort taking Junior Certificate mathematics in 2003, it is 82.15.[36] Table A4.17 shows the Pearson correlations between achievement on the PISA mathematics scales and the eight MJCPS scales explored for the sub-cohort. The pattern of correlations is also very similar to that observed for all students participating in PISA 2000 (Table 4.20), although the correlations are marginally higher. The same pattern was observed for the English/reading analyses of the sub-cohort of students taking the Junior Certificate in 2000.

4.5.5. Confirmatory Analyses Using the PISA 2000 Mathematics / Junior Certificate Mathematics Data

To account for differences in the PISA 2000 and PISA 2003 samples that might have arisen from sampling fluctuation or other sources, mean mathematics scores associated with MJCPS scale 3 (Table 4.15) (which is the same as the final scale, scale 8, with the exception of having a 2-point gap between the lowest and second-lowest groups, and the highest and second-highest) were computed for the PISA 2000 cohort – both the overall sample, and the sub-cohort attempting the Junior Certificate mathematics examination in 2000. Means for the overall scale, and for the Space & Shape and

[36] These standard deviations are the weighted averages of the individual standard deviations associated with the five plausible values for combined mathematics.

Change & Relationships subscales only were computed.[37] Pearson correlations associated with these were also computed. Results are shown in Table A4.18, A4.19 and A4.20. The Ns are smaller than the total sample N since PISA 2000 mathematics scores are available only for those students who attempted mathematics (2128; five in nine students) while in PISA 2003 achievement scores are available for all students and for all domains. Comparing the magnitude of the Pearson correlation coefficient for scale 3 (Table A4.18) (.700 for all PISA 2000 students for the combined mathematics scale), it is somewhat lower than the correlation reported for the original 12-point MJCPS (.742; Shiel et al., Table 6.14). The correlations for the subscales are even lower (.450 for Space & Shape and .505 for Change & Relationships), but this may relate to the small numbers of items making up each of the subscales. A comparison of the mean scores for each scale point for the Space & Shape and Change & Relationships subscales reveal some interesting differences (Tables A4.19 and A4.20). In PISA 2000, students at the lower end of the MJCPS did better when compared to the equivalent points in 2003, while those at the upper end did worse. In other words, the distribution of mean scores on the Space & Shape and Change & Relationships scores across MJCPS scale points is more homogenous in 2000 than in 2003. When one compares the mean scores at each MJCPS scale point for the combined scale, the pattern is somewhat more consistent, although there are some differences. For example, the score difference between the lowest and second lowest points on the scale for the PISA 2000 sample (about 56 points) is larger than that for the PISA 2003 sample (about 41 points). In contrast, the score difference between the highest and second highest points on the scale for the PISA 2000 sample (about 32 scale points) is smaller than that for the PISA 2003 sample (about 46 scale points). As with the differences observed for English/reading when comparing the two years, these fluctuations may reflect differences across the two years in performance amongst the lowest and highest achievers, differences in marking the Junior Certificate mathematics examination (e.g., perhaps a grade A at higher level was more difficult to attain in 2003 than in 2000), or a mixture of these. The small numbers of students at the lower end of the scale and the large standard errors associated with the

[37] In PISA 2000, there were no mathematics items measuring Quantity and Uncertainty. It should further be noted that item parameters for the PISA 2000 combined mathematics scale were subject to a 'booklet ordering' effect since the rotated booklet design was not balanced. The 2003 booklet design, however, was balanced, so the 2003 item parameters were re-applied to the PISA 2000 achievement data to produce the two subscales (OECD, 2005). Thus in terms of comparing across 2000 and 2003, the mathematics subscales rather than the combined scales should be used.

lower end also contribute to this. It could be inferred that the manner in which MJCPS might 'best' be scaled depends upon the PISA cycle in question.

4.5.6. The Preferred MJCPS

Due to the dips in the mid-upper regions of scales 5 and 6 and the overlap observed in scale 7 between MJCPS groups 9 and 10, it was decided to select scale 8 as the preferred scale for the MJCPS. This is a 10-point scale with a one-grade overlap between ordinary and foundation, a 3-grade overlap between higher and ordinary, and two scale points between the lowest and second lowest groups, and between the highest and second highest groups.

As with the EJCPS selected, foundation-level students obtaining grades B, C, D, E and F, and ordinary-level students obtaining grades E and F (321 students), cannot be distinguished from one another on the PISA mathematics scales, and this group of students forms a single group of low achievers at the lower tail of the MJCPS. The analyses provide further support for the 3-grade overlap first suggested by Martin and Hickey (1992) and used by Shiel et al. (2001) and Cosgrove et al. (2005), at least between higher and ordinary level. There is also some evidence of stretching at the very upper end of the scale, with a larger than average mean score difference between students scoring an A and a B at higher level.

4.6. Conclusion

The aim of this chapter was to enhance understanding of the relationship between performance on PISA and the Junior Certificate. This task is perhaps more straightforward in the case of English/reading where congruence between the reading processes that PISA and the Junior Certificate appear to assess was noted in Chapter 2. In the case of mathematics, however, since the two assessments differ widely, interpretation is not so straightforward. One theme of interest in this regard is whether and to what extent performance differences across the four mathematics subscales are in evidence and whether these can be interpreted with respect to curricular content. Further, it was noted in Chapter 2 that while different ways of scaling the Junior Certificate had been explored with the PISA 2000 dataset, these analyses did not consider whether the Junior Certificate grades are stretched at the upper end and/or more clustered at the lower end, as one might expect from public examinations data

(Greaney & Kellaghan, 1996; Millar & Kelly, 1999). The availability of the PISA test data as an independent achievement measure (again, perhaps not an ideal one in the case of mathematics) allows one to explore these two possibilities. Overall, the analyses are intended to add to the research on curriculum and PISA reviewed in Chapter 2.

Analyses looked initially at the distribution of achievement by Junior Certificate syllabus level for English in 2000 and mathematics in 2003, and then (taking measurement and sampling error into account) looked at various possible ways to scale these two Junior Certificate subjects. The focus was initially on finding the best way to scale the lower end of the JCPS (e.g., whether some grades at ordinary and foundation levels should be collapsed into a single group); then on examining the best amount of overlap of grades between syllabus levels (e.g., whether a smoother scale might be produced with a two-grade overlap between higher and ordinary levels compared with a three-grade overlap); and finally, on examining whether it would be necessary to stretch the JCPS at the extremes in order to create a scale with approximately equal achievement intervals between scale points (e.g., if obtaining an A grade at higher level was associated with particularly high performance on PISA, then it should appear more than one point on the scale above a B grade). To address the possibility that links between the two scales may vary depending on the year in which the Junior Certificate was taken, on the status of the PISA domain as major or minor, or other fluctuations, analyses for both subject areas were carried out for both 2000 and 2003, as were analyses of the sub-cohort within each PISA cycle taking the Junior Certificate Examination in the same year as PISA.

Analyses comparing the mean PISA scores of students at each point on a number of possible JCPS for English/reading and mathematics suggest that a 9-point scale for Junior Certificate English is the most appropriate, given the available data. This scale is similar to the original 12-point EJCPS used in analyses in Shiel et al. (2001) and Cosgrove et al. (2005) in that there is a 3-grade overlap between higher and ordinary levels (such that an A at ordinary level is considered equivalent to a D at higher level). However, students attaining a grade E or F at ordinary level, and students attaining below grade A at foundation level, cannot be distinguished from one another in terms of their achievement on the PISA reading scales and so the lowest five points on the original EJCPS were collapsed into a single point, providing support for the argument

that the scale is more clustered at the lower end; i.e., the achievements of many ordinary- and foundation-level students on PISA are indistinguishable. A 2-point interval was introduced between the lowest and second lowest point in order to have an EJCPS of roughly equal intervals. The final scale has 9 points with an average of 36 to 39 PISA scale points between EJCPS points (depending on the PISA reading scale/subscale considered).

Average performance across the reading subscales at each EJCPS scale point mirrors overall average performance in the case of the Retrieve and Interpret subscales, but comparatively strong performance of the lowest achievers on the Reflect subscale is evident. Students scoring a 1 on the EJCPS achieved a mean score of 339 on the combined reading scale, and scores of 329, 333, and 355 on the Retrieve, Interpret and Reflect subscales, respectively. In contrast, students scoring at the highest point on EJCPS achieved mean scores of between 639 and 645 on the reading scales; i.e., there was less variability across the subscales at the upper end of the EJCPS distribution. This pattern holds for the sub-cohort taking the Junior Certificate in 2000 only and suggests that Ireland's particularly strong average performance on the Reflect subscale is attributable to strong performance on Reflect items by lower achievers. Why this is so is impossible to say. It may be the case that lower achievers were more motivated to attempt these items (e.g., found them to be more engaging), or, given that proportionately more Reflect items required extended written responses compared with Retrieve and Interpret items, were more familiar with the item format.

There was no evidence of stretching at the upper end of the EJCPS (e.g., the PISA score point difference between students obtaining an A and a B at higher level is similar to the score point difference between B and C): PISA score differences between the upper and middle regions of the scale are similar. Analyses of the sub-cohort of PISA 2000 students taking the Junior Certificate English examination in 2000 also provided support for the choice of the 9-point EJCPS. The only notable difference in PISA mean scores at various points on the EJCPS is that scores are about one-sixth of a standard deviation lower than the sample as a whole. This score difference can be at least partly attributed to the fact that students taking the Junior Certificate in 1999 have had one year's extra schooling compared to those taking it in 2000 at the time of taking the PISA test. Mean scores on the PISA 2003 reading scale at each point of the 9-point EJCPS were also

computed for the PISA 2003 sample and compared to the mean scores of the 2000 sample. Some fluctuations were observed: those at the lower end of EJCPS in 2003 did better than those in 2000, whereas the reverse is the case for students at the upper end of the EJCPS. This suggests that achievement estimates at the extremes of the distribution may be inherently less stable than those at the middle. There are low numbers of students at the lower end of the scale and there are large standard errors associated with their mean scores. Further, fewer reading items, and with a narrower difficulty range, were used in PISA 2003 compared to 2000. In PISA 2003, there were 28 reading items with scaled item difficulties ranging from 336 to 774. In PISA 2000, there were 141 items with scaled item difficulties ranging from 341 to 822 (Adams & Wu, 2002; OECD, 2005b).

In the case of mathematics, the preferred MJCPS identified in the analyses is identical to the EJCPS scale selected, except that there is a 2-point interval between the second highest and highest scale points, as well as between the second lowest and lowest points. Hence, there is some evidence of both stretching at the upper end of the scale and clustering at the lower end. However, the absolute difference between the uppermost and lowest points on the MJCPS – 231 points – is smaller than the absolute difference on the EJCPS (304 points), which is consistent with the comparatively small standard deviation on PISA mathematics relative to PISA reading. A comparison of the mean scores on each of the mathematics subscales at each point on the MJCPS indicates that performance across the MJCPS distribution mirrors overall average performance; so while one can speculate about the reasons for the stronger than expected performance of lower achievers on the Reflect subscale, no such fluctuations are evident for mathematics. Analyses of the sub-cohort of PISA 2003 students taking the Junior Certificate mathematics examination in 2003 also provided support for the choice of the 10-point MJCPS. The only notable difference in PISA mean scores at various points on the MJCPS was that scores were about one-eighth of a standard deviation lower than for the sample as a whole (a pattern also observed in the EJCPS sub-cohort).

Mean scores on the PISA 2000 mathematics scale at each point of the 10-point MJCPS were computed for the PISA 2000 sample for the Space & Shape and Change & Relationships scales and compared to the mean scores of the 2003 sample. For both subscales, the mean scores at the lower end of the 2000 MJCPS were higher than those

of the 2003 sample; those at the upper end of the 2000 MJCPS were lower. Again, as with English/reading, reasons for this are not clear; the small numbers of items associated with the subscales in 2000 may have been a factor. Just 10 items contributed to each of the subscales in PISA 2000 and these had scaled item difficulties ranging from 420 to 723. In contrast, there were 22 items associated with each of these subscales in 2003 and these had scaled item difficulties ranging from 262 to 801 (Adams & Wu, 2002; OECD, 2005b). Indeed, when the combined scales for mathematics for 2000 and 2003 were compared, results were more similar.

It is regrettable that the Junior Certificate Examinations database does not preserve the 'raw' percent score for individual Junior Certificate subjects. Converting percent scores to a 6-point scale (A to F) results in a loss of much information that would have been likely to have been of use in attempts to scale achievements on Junior Certificate English and mathematics.

In conclusion, the inability of the PISA achievement measure to distinguish between students achieving grades E or F at ordinary level, and below grade A at foundation level, should be considered in the light of the differential response rates of low and high achievers in general (described in Chapter 3), resulting in lower numbers at the lower end of the JCPS than would have been observed in the population. It may be the case that, had more ordinary- and foundation-level students participated in PISA, the lower points of the JCPS might have been empirically distinct.

However, this finding could also have been due in part to the limited ability of *both* assessments to discriminate between the achievements of students at the lower ends of the achievement distribution. It was noted that there are relatively few PISA items in the reading and mathematics items pools which assess knowledge and skills at this point. It was also noted in Chapter 2 that the marking schemes for the Junior Certificate appear to provide opportunity for merit on some of the examination questions, which may result in higher 'pass' rates of the lower achievers, although no detailed research on this possibility has been undertaken.

Notwithstanding non-response bias and large measurement errors, the fact that the performance of many students at ordinary and foundation levels is indistinguishable

calls into question the appropriateness of the ordinary-level course for some students. The very low performance on both PISA reading and mathematics of some ordinary-level students was noted in Chapter 2. This is notable particularly the absence of concrete guidelines to schools and teachers as to which syllabus levels might be best suited to which students. The lack of research into how and why students come to take the Junior Certificate Examinations at particular syllabus levels was also noted in Chapter 2.

These findings also have implications for PISA, both what it measures and its survey design. They suggest that the PISA achievement measures of both reading and mathematics are of very limited use in describing the achievements of foundation-level students and students at the lower end of the letter grades at ordinary level since the majority of these students are 'off the scale'. The limited utility of PISA in this regard is compounded by the differential student non-response observed in Chapter 3, and how this is treated in the weighting process. This is quite a serious shortcoming considering the reliance on the percentages of students at or below Level 1 as a key indicator of the performance of the education system. A more accurate measure of performance at the lower end of the scale would seem highly desirable, particularly for monitoring trends across time (and if this indicator proves, as Coulombe et al., 2004, have suggested, to be a key measure of economic competitiveness).

CHAPTER 5. A COMPARISON OF ACHIEVEMENT VARIANCE AND EXPLANATORY MODELS OF PISA, THE JUNIOR CERTIFICATE AND TIMSS

5.1. Introduction

The analyses presented in this chapter attempt to build on the review of variance components and explanatory models of achievement presented in Chapter 2 and aim to extend understanding about what PISA can tell us about the equity of achievement outcomes and the determinants of achievement. Both are considered together in this chapter since the same analytic framework underpins them, i.e., multilevel modelling (or hierarchical linear modelling). It was noted in Chapters 1 and 2 that three aspects of survey designs in particular impact on these two issues: sample design (whether age- or grade-based), curriculum sensitivity of the test measure, and school dependence/independence of the test measure. Ideally, these analyses should be conducted using the dataset from a single survey (e.g., if PISA had incorporated a hybrid age-based sample design plus the sampling of intact third year classes, it would be possible to make direct, straightforward comparisons) but this is not possible. Therefore, the best available data are used to explore these three issues and to provide initial evidence for the arguments made.

5.2. Rationale

The importance ascribed to the between-school variance statistic as a measure of educational equity was noted in Chapter 1. However, one might draw very different conclusions about the equity of Ireland's education system depending on the sample design. There may also be variations depending on whether the test measure is intended to be curriculum-sensitive or not, and whether the subject domain is school-dependent or more generic, although the research reviewed in Chapter 2 which directly addresses these issues is almost 30 years old and changes in the education system (such as increased rates of enrolment and revisions to curricula) make it impossible to say whether these findings still hold. However, given that the between-school variance components for TIMSS 1995 and PISA 2000/2003 are very different for Ireland (with the former being much higher), that the published variance components for TIMSS are based on a single intact class per school (despite the fact that two classes per school were selected), and that explanatory models of Irish achievement reviewed in Chapter 2

indicate within-school selection according to ability, the TIMSS data are re-visited in a comparison of the published variance components for mathematics from TIMSS 1995 and PISA 2000 for countries participating in both studies. Then, the variance components for TIMSS 1995 for Ireland are re-computed using all available data (i.e., two intact classes per school) and compared with the original estimates. These are also compared with the variance components for performance on Junior Certificate mathematics to investigate whether the more curriculum-sensitive measure is associated with higher between-cluster variance. The variance components for PISA 2000 reading, Junior Certificate English, and PISA 2003 mathematics and Junior Certificate mathematics are also considered with respect to the variance components for TIMSS. Also, the variance components for third year students only for the PISA datasets are compared with the full datasets to investigate whether between-school variance increases if one constrains the sample to a single grade level.

In considering the determinants of achievement, six multilevel explanatory models of achievement are presented: PISA 2000 reading, 2000 Junior Certificate English, PISA 2003 mathematics, 2003 Junior Certificate mathematics, TIMSS 1995 mathematics and 1996 Junior Certificate mathematics. As mentioned already, comparisons across TIMSS, PISA and the Junior Certificate are complex since the TIMSS and PISA surveys and achievement measures differ in several important respects, notably the population surveyed, the extent to which the achievement measure diverges from the Junior Certificate mathematics syllabus/examinations, and the sample design. These, however, are the best available data with which to investigate the issues outlined. Insofar as possible, the models have been constructed so as to maximise the comparability of results. In the case of the PISA models, grade 9/third year students only are included. This reduces the complexity of interpreting results which pertain to four grade levels. Further, the variables in the models have been selected to be broadly comparable with one another, although the items comprising the composite measures and the methods used to construct them differ somewhat across TIMSS 1995 and PISA 2000/2003.

Three major themes are explored in these models: the nature of the test measure, the nature of the sample design, and the strength of the impact of social intake and school/class variables on achievement (see Table 2.14). Several hypotheses are

explored. Regarding the curriculum sensitivity and school-dependent nature of the test, it is hypothesised that, if mathematics is more school-dependent than English/reading, mathematics achievement will be more sensitive to school-level effects than measures of English/reading. (In the present discussion, 'school-level effects' refers to associations between achievement and school/class variables *other than* school-level socioeconomic status.) Second, because the test-curriculum rating project described in Chapter 2 suggested notable disparities between PISA mathematics and Junior Certificate mathematics, both in the concepts assessed and in the manner in which problems are contextualised, and because TIMSS mathematics is intended to be only somewhat compatible with national mathematics curricula, it is hypothesised that Junior Certificate mathematics will be more sensitive than *both* PISA mathematics and TIMSS mathematics to school-level effects. Third, it is hypothesised that PISA mathematics will be least sensitive to such effects. Fourth, because the test-curriculum rating project suggested similarities in the reading processes assessed in PISA reading and Junior Certificate English, it is hypothesised that the explanatory models for both of these will be highly similar, assuming that the domain of English/reading is less school-dependent and that familiarity with the reading process assessed is more relevant to success on these assessments than the item format or type or length of text.

Regarding the impact of social intake, it is hypothesised first, that the association between school-level SES and achievement will be strong in all models examined. Second, it is hypothesised that the social context effect will be somewhat weaker in the models of mathematics compared with English/reading (cf. Sofroniou et al., in preparation). Third, whether or not the above two hypotheses also hold across measures (whether curriculum-sensitive or not) will be investigated. Fourth, if students are clustered within classrooms on the basis of social background, as well as on the basis of ability, the strength of the effect for social intake will be stronger for the models in which samples are based on intact-class sampling will be stronger than for models in which samples were drawn at random within schools.

The models will also investigate whether the association between social intake and achievement is linear or curvilinear; and also whether there is an interaction between student gender and social intake. Given the findings of Sofroniou et al. (in preparation) in particular, it is expected that the social intake effect will be linear and that in at least

some of the models, there will be an interaction between student gender and school SES (whereby the effect is stronger for males). These are explored since the use of the PISA results to inform policy on educational disadvantage in Ireland has been mentioned in ministerial commentary on the results (see Chapter 1); clarifying the nature of the relationship of social background with achievement, and whether it differs for boys and girls, may enhance our understanding and add to policy development.

5.3. Impact of Sample Design and Test Content on Between-School Variance in Achievement

In Chapter 2, it was found, in a comparison of the variance components of TIMSS 1995 and PISA 2000, that higher between-cluster variance was associated with treating the class, rather than the school, as the unit of analysis. It was also noted in a comparison of the variance components associated with achievements on PISA 2000 reading and students taking the Junior Certificate English examination in 1999 or 2000 that there was little difference between the two measures in the between-school variance. However, a comparison of PISA 2000 mathematics and Junior Certificate mathematics suggests that between-school variance is slightly higher for the Junior Certificate measure.

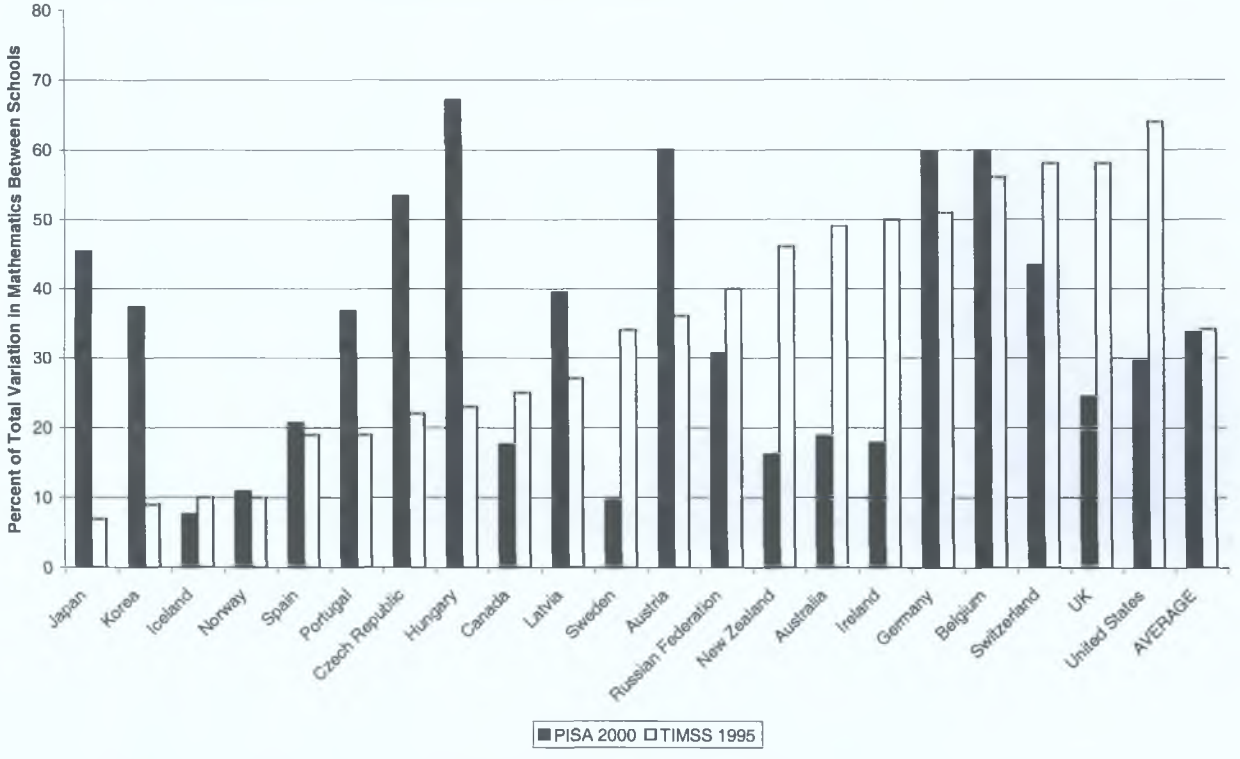
It is not clear, however, to what extent the sample design confounds the interpretation of variance components in the case of curriculum-sensitive and more generic test measures. In this section, more detailed comparisons of the between-cluster variance associated with students who participated in TIMSS 1995, PISA 2000 and PISA 2003 are made in an attempt to clarify this issue.

5.3.1. A Re-analysis of the Variance Components of TIMSS and PISA: International Comparisons

A comparison of the between-school variance in mathematics achievement for countries participating in both TIMSS 1995 and PISA 2000 reveal large differences in the two measures. Figure 5.1 compares the percentage of total variance in mathematics achievement that is between schools in TIMSS 1995 and PISA 2000 for the 21 countries that participated in both surveys and for whom data are available. Countries are ranked in ascending order of the value for TIMSS. This comparison shows first, that the average between-school variance of these countries is almost identical for the two

surveys – in the region of 33%. Second, notwithstanding the fact that TIMSS attempted to measure aspects of the curriculum in participating countries rather than having its basis in 'real-life literacy', as well as the five-year gap between surveys, the conclusions one might draw about which countries demonstrate homogeneity in achievement outcomes differs substantially, depending on which survey is considered.

Figure 5.1. A Comparison of the Percentage of Variance in Mathematics that is Between Schools: PISA 2000 and TIMSS 1995



Note. Countries are ranked in ascending order by percent of variance between schools on TIMSS.

On the TIMSS measure, a large between-school difference in achievement in Ireland is evident, comparable to New Zealand, Australia, Germany and Belgium. On PISA, the picture is very different. Ireland has comparatively low between-school variance, similar to Spain, Canada, New Zealand and Australia. The pattern for Ireland for both TIMSS and PISA is similar to the pattern observed for Sweden, New Zealand, Australia, and the UK. One could hypothesise that in these five countries, considerable achievement differences are occurring *within* schools due to ability streaming, or other reasons relating to curriculum content/delivery, while selection at the level of schools is not strongly associated with variables related to achievement. It is unfortunate that Martin et al. (2000b) based country comparisons of between-school variance on grade 8

students only, since this is likely to confound between-school and between-class variance.

5.3.2. National Comparisons of Variance Components

5.3.2.1. Procedure

Comparisons were made using HLM 6.0 (Raudenbush, et al., 2004). Variance components associated with a series of null hierarchical linear models were computed. The so-called null model includes only the outcome measure, and no explanatory variables, and allows one to partition the total variance in achievement into different components which reflect the clustered nature of the data (e.g., variation between students, classes and schools). The three-level null model partitions the variance of the outcome variable into between-school, between-class, and between-student components. Two-level models can partition the outcome variable into between-student and between-school or between-class components (depending on the sample design). (Section 5.4 describes some key concepts associated with hierarchical linear models in more detail.)

5.3.2.2. Comparison of TIMSS 1995 Mathematics and Junior Certificate Mathematics

Four models for each of TIMSS mathematics and of Junior Certificate Performance Scale scores in mathematics (MJCPS) for students participating in TIMSS and who took the Junior Certificate Examination in mathematics in 1996 or 1997 (i.e., one or two years after the TIMSS assessment) are compared in Table 5.1. MJCPS scores are available for mathematics for 94.1% of TIMSS students.[38] The three-level model partitions variance into between-school, between-class, and between-student components. The first two-level model takes the class as the cluster variable for all participating students; the second takes the school as the cluster variable for all participating students; the third looks at variance for grade 8 (second year) students only. The results of this exercise confirm the confounding of between-class and between-school variance in the published results for TIMSS (Martin et al., 2000b).

[38] Total N for Irish TIMSS participants = 6203; of these, 5834 have MJCPS scores.

Table 5.1. TIMSS 1995 Mathematics and Junior Certificate 1996/1997 Mathematics: Variance Components for Ireland: Comparison of Variance Components of 3-Level and Various 2-Level Null Models

3 Level Model

	TIMSS		*Junior Cert.*	
	Variance	% of total	Variance	% of total
Between students	5127.02	61.9	2.406	48.1
Between classes	2371.53	28.6	1.703	34.0
Between schools	788.18	9.5	0.893	17.9
Total variance	8286.73	100.0	5.002	100.0

2 Level Model, all students, with class as the cluster variable

	TIMSS		*Junior Cert.*	
	Variance	% of total	Variance	% of total
Between students	5127.03	61.8	2.406	48.0
Between classes	3171.06	38.2	2.606	52.0
Total variance	8298.09	100.0	5.012	100.0

2 Level Model, all students, with school as the cluster variable

	TIMSS		*Junior Cert.*	
	Variance	% of total	Variance	% of total
Between students	6279.20	76.0	3.190	65.5
Between schools	1982.25	24.0	1.680	34.5
Total variance	8261.46	100.0	4.869	100.0

2 Level Model, grade 8 students only

	TIMSS		*Junior Cert.*	
	Variance	% of total	Variance	% of total
Between students	4897.40	55.4	1.827	36.4
Between classes/schools	3941.86	44.6	3.194	63.6
Total variance	8839.27	100.0	5.022	100.0

Note. Analyses of the TIMSS data used all 5 plausible values; all analyses are unweighted.

In the case of the TIMSS achievement data, the three-level model shows, consistent with Madaus et al. (1979), that proportionately more of the achievement variance is between classes compared with schools. The between-student variance for TIMSS mathematics is 61.9%. Of the remainder, 28.6% is between classes and 9.5% between schools. The first and second two-level models of TIMSS in Table 5.1 confirm that much of the within-school variance is between classes. If students are clustered by class, the between-cluster variance for TIMSS mathematics is 38.2%; if by school, it is 24.0%. The third two-level model for TIMSS mathematics shows that the between-school or between-class variance is inflated to almost 45% if only the upper grade is included.[39]

[39] This figure is somewhat lower than those reported by Martin et al. (2000b), but the present analyses used unweighted data and all five plausible values, with full maximum likelihood estimation in HLM 6.0, while Martin et al. do not give details about the method they used.

The same broad pattern holds for Junior Certificate mathematics achievement data of students that participated in TIMSS; i.e., the three-level model shows that proportionately more variance in achievement is between classes than between schools; that treating the class as the cluster variable results in a higher between-cluster variance than if the school is the cluster variable, and that the selection of grade 8 students only in the computation of between-'school' variance results in the highest between-cluster variance of all models examined.

A comparison of TIMSS 1995 mathematics with the Junior Certificate mathematics variance components indicates that the Junior Certificate measure is comparatively more sensitive to school/class effects (again, consistent with Madaus et al., 1979). In the three-level model for TIMSS mathematics for example, the between-class and between-school variances are 28.6% and 9.5%, respectively. The values for MJCPS are *both* higher, at 34.0% and 17.9%, respectively. The between-school variance for Junior Certificate mathematics when grade 8 students are considered separately is close to two-thirds of the total variance (63.6%), and again, higher than the corresponding value for TIMSS (44.6%).

5.3.2.3. Comparison of PISA 2000 Reading with Junior Certificate English, and of PISA 2003 Mathematics with Junior Certificate Mathematics

Table 5.2 compares the variance components for six two-level models of English/reading: PISA 2000 reading, Junior Certificate English 12-point performance scale, and Junior Certificate English 9-point performance scale described in Chapter 4, for the PISA 2000 sample as a whole and also for the subset of students attempting the Junior Certificate in 2000 (i.e., grade 9/third year students only). Results indicate that there are only small differences in the proportions of variance between schools, regardless of which measure is considered, and whether or not one excludes students who are not in grade 9. Between-school variance is marginally higher – by about 2% – for the Junior Certificate measures. Between-school variance is also marginally higher – again by about 2% – for the subset of students attempting the Junior Certificate in 2000.

Table 5.2. Variance Components for Ireland - Students Participating in PISA 2000 and Who Took Junior Certificate English in 1999 or 2000: Comparison of Various 2-Level Null Models

All students with EJCPS Data for 1999 or 2000						
	PISA Reading		*EJCPS (12-point)*		*EJCPS (preferred 9-point)*	
	Variance	% of total	Variance	% of total	Variance	% of total
Between students	1402.20	16.9	0.553	18.8	0.559	18.9
Between schools	6872.71	83.1	2.385	81.2	2.403	81.1
Total variance	8274.91	100.0	2.938	100.0	2.962	100.0
Students with EJCPS Data for 2000 only						
	PISA Reading		*EJCPS (12-point)*		*EJCPS (preferred 9-point)*	
	Variance	% of total	Variance	% of total	Variance	% of total
Between students	1569.52	18.6	0.632	20.5	0.639	20.6
Between schools	6891.38	81.4	2.448	79.5	2.462	79.4
Total variance	8460.91	100.0	3.080	100.0	3.101	100.0

Note. Analyses of the PISA data used all 5 plausible values; all analyses are unweighted.

Table 5.3 compares the variance components for six two-level models of mathematics: PISA 2003 mathematics, Junior Certificate mathematics 12-point performance scale, and Junior Certificate mathematics preferred 10-point performance scale (described in Chapter 4), for the PISA 2003 sample as a whole and also for the subset of students attempting the Junior Certificate in 2003. Results indicate that there is no appreciable difference in the between-school variance of the subset of students attempting the Junior Certificate in 2003 compared with the sample as a whole. However, the between-school variance associated with Junior Certificate mathematics, although low overall (around 20%) is consistently higher than the between-school variance associated with PISA mathematics (around 15%). This finding is consistent with a comparison of variance components for PISA 2000 mathematics and reading (Shiel et al., 2001) and students taking Junior Certificate mathematics and English in 1999 or 2000 (Sofroniou et al., 2000; 2002), which indicates that there is no difference in the between-school variances associated with PISA reading and Junior Certificate English, but that the between-school variance for Junior Certificate mathematics is a little higher than PISA mathematics. Hence, there is tentative evidence to suggest that Junior Certificate mathematics may be more sensitive than the international assessment measures to school/class effects than both Junior Certificate English and PISA mathematics.

Table 5.3. Variance Components for Ireland - Students Participating in PISA 2003 and Who Took Junior Certificate Mathematics in 2002 or 2003: Comparison of Various 2-Level Null Models

All students with MJCPS Data for 2002 or 2003						
	PISA Mathematics		*MJCPS (12-point)*		*MJCPS (preferred 10-point)*	
	Variance	% of total	Variance	% of total	Variance	% of total
Between students	1095.97	15.6	0.902	19.8	1.072	19.4
Between schools	5908.78	84.4	3.652	80.2	4.448	80.6
Total variance	7004.76	100.0	4.554	100.0	5.520	100.0
Students with MJCPS Data for 2003 only						
	PISA Mathematics		*MJCPS (12-point)*		*MJCPS (preferred 10-point)*	
	Variance	% of total	Variance	% of total	Variance	% of total
Between students	998.72	14.8	0.975	20.8	1.162	20.2
Between schools	5763.10	85.2	3.716	79.2	4.585	79.8
Total variance	6761.82	100.0	4.691	100.0	5.747	100.0

Note. Analyses of the PISA data used all 5 plausible values; all analyses are unweighted.

Tables 5.2 and 5.3 also indicate that there is minimal, if any, consequence, in the interpretation of variance components whether one uses the preferred Junior Certificate scales (described in Chapter 4) compared with the original 12-point ones. Assuming that any changes in the Junior Certificate mathematics curriculum and its assessment between 1995 and 2003 do not have a substantial impact on between-school variance, a comparison of the variance components for Junior Certificate mathematics in Table 5.1 and 5.3 suggests that the sample design and method of selecting students (intact classes versus random sample across multiple grade levels/classes) exerts a substantial influence on the interpretation of between-school variance.

5.4. Key Concepts Associated with Multilevel Models

Prior to presenting the analyses, a brief description of some of the main concepts associated with multilevel models is given. Multilevel models (also referred to as hierarchical linear models or mixed models) provide a flexible approach to the analysis of non-independent or 'clustered' data that arise when studying topics such as students nested within classrooms. Traditional general linear models (e.g., ordinary least squares regression) are not well-suited for the analysis of these types of data, given the violation of the assumption of independence, and disaggregating cluster-level data to the level of the individual results in an under-estimate of standard errors associated with estimates of cluster-level variables (Osborne, 2000). In contrast, multilevel models are explicitly designed to analyse clustered data structures and can incorporate individual-level

predictors, group-level predictors, and individual-by-group-level interactions (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999).

The most basic form of the two-level random intercepts model may be written as

$$y_{ij} = \beta_j + e_{ij}$$
$$= \beta_0 + u_j + e_{ij}$$

Where y_{ij} is the outcome of student i in school j; this is expressed as a school component, β_j and an error component, e_{ij}, which is the difference between each individual student's score and the average score for the school s/he is in. The school component is expressed as the school mean β_0 and the deviation from each school's mean from the overall mean (u_j). The school error term is the distinguishing feature of a multilevel model compared to an ordinary-least-squares regression model. In addition to the error term at the school level, estimates are multiplied by a shrinkage factor, whereby the estimated school mean is closer to the overall mean in cases where there are few students in a school (Goldstein, 1997). The two random terms can be summarised by their variances, δ_e^2 and δ_{u0}^2. The proportion of variance between clusters (the intra-cluster correlation) is defined as

$$\delta_{u0}^2 / (\delta_e^2 + \delta_{u0}^2)$$

The smaller the intra-cluster correlation, the less schools vary with respect to achievement. The random intercepts model can be extended to incorporate explanatory variables at the student level to explain between-and within-cluster achievement, e.g.,

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_j + e_{ij}$$

Where x_1 is the gender of student i in school j; x_2 is the socioeconomic status of student i in school j. The model can be extended further to include explanatory terms at the school level, e.g.,

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3j} + u_j + e_{ij}$$
$$= \beta_0 + \beta_3 x_j + u_j + e_{ij}$$

Where x_3 is the average socioeconomic status of the students in school j. A random term may be added to the slope of each variable as well as the intercept for a student-level variables to allow its effects to vary across schools – the fully random model (or 'means as outcomes' model; Raudenbush & Bryk, 1992), e.g.,

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3j} + u_j + e_{ij}$$
$$\beta_{1j} = \beta_1 + u_{1j}$$

In this example, the effect of student gender on achievement varies across schools. Interaction terms within a level may be tested in a similar fashion to interactions in ordinary-least-squares regression, e.g., the interaction between gender and socioeconomic status is expressed as follows

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{1ij} x_{2ij} + u_j + e_{ij}$$

A third type of interaction may occur, also termed 'slopes as outcomes' (Preacher, et al., 2003); i.e., where the slope of a student-level variable interacts with school-level variables, a so-called cross-level interaction. An example of this is an interaction between student gender and school socioeconomic status (SES). This type of interaction is expressed as follows within a multilevel modelling framework (where u_{1j} is the error associated with the slope for gender, and β_{4j} is school SES):

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3j} + u_j + e_{ij}$$

$$\beta_{1j} = \beta_{4j} + u_{1j}$$

5.5. Questions Addressed in the Analyses

Based on the literature reviewed in Chapters 1 and 2, the analyses presented in this chapter explore the following research questions, which were elaborated on in Section 5.2.

1. *Test Content:* Taking into account the curriculum sensitivity of the test, as well as the extent to which it may be considered school-dependent or school-independent, it is hypothesised that
    - Mathematics achievement will be more sensitive to school/class effects (other than SES) than measures of English/reading.
    - Junior Certificate mathematics will be more sensitive than *both* PISA mathematics and TIMSS mathematics to school-level effects and PISA mathematics will be least sensitive to such effects.
    - The explanatory models for PISA reading and Junior Certificate English of these will be highly similar to one another.
2. *Social intake:* Given the results of explanatory models reviewed in Chapter 2, it is hypothesised that
    - the association between school-level SES and achievement will be strong in all models examined.

- the effect of social intake will be somewhat weaker in the models of mathematics compared with English/reading (cf. Sofroniou et al., in preparation).
- whether or not the above two hypotheses also hold across measures (whether curriculum-sensitive or not) will be investigated.
- if students are clustered within classrooms on the basis of social background, as well as on the basis of ability, the strength of the effect associated with social intake will be stronger in models involving students who participated in TIMSS compared with models involving students who participated in PISA.

5.6. Procedure

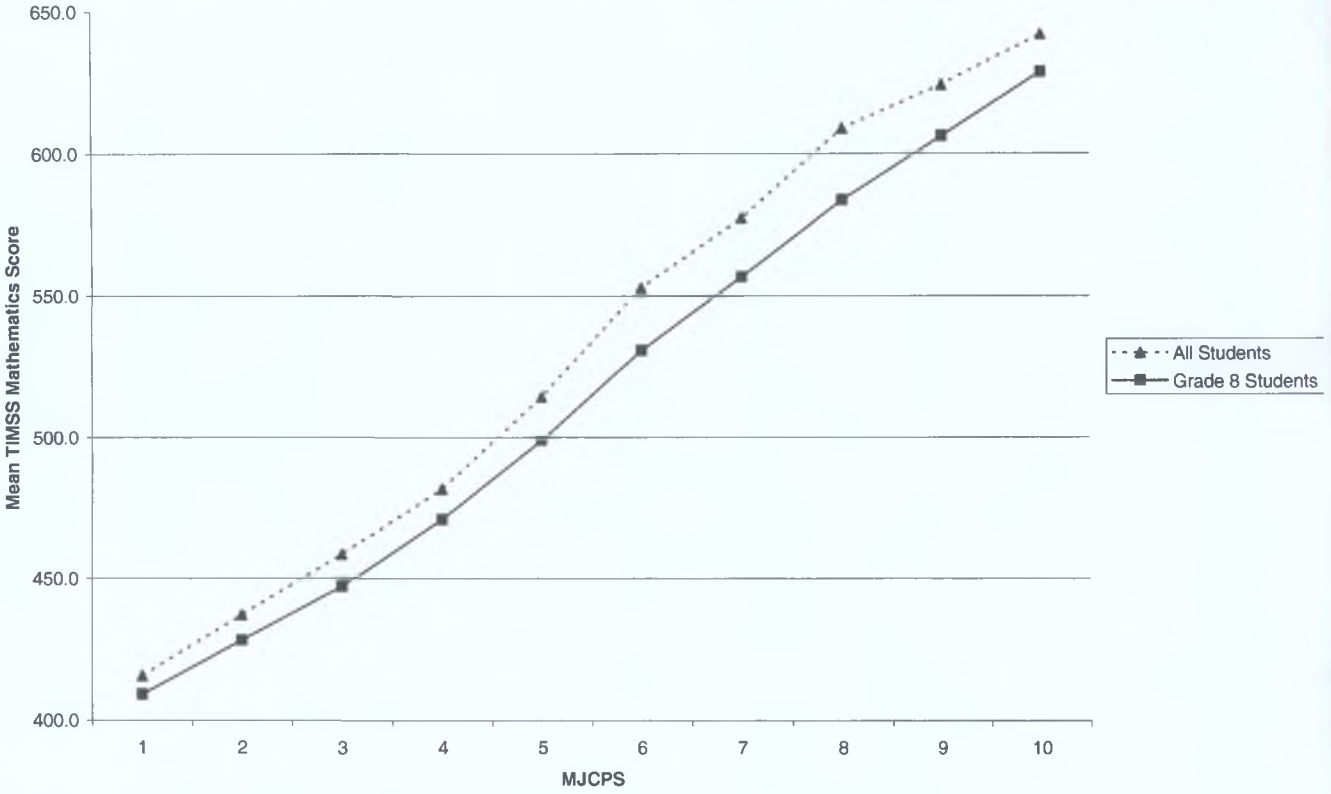
5.6.1. Constructing the models

I compare results of a series of multilevel models (hierarchical linear models) for students participating in PISA 2000, PISA 2003, and TIMSS 1995. Rather than including all students who participated in the survey in question, it was decided to examine the achievements only of students who took the Junior Certificate in the same year as the PISA assessment and, in the case of TIMSS, only students who took the Junior Certificate in 1996. The data in Tables 5.2 and 5.3 indicate that the selection of grade 9 students has little if any effect on the proportion of variance that is between schools for students participating in PISA 2000 and PISA 2003. In the case of TIMSS, however, the selection of grade 8 students results in an increase in the between-school variance since each school sample corresponds to one intact class (Table 5.1).

In all models of the Junior Certificate, the preferred Junior Certificate scales described in Chapter 4 are used as the outcome measure. However, the question arises as to whether this method of scaling the Junior Certificate mathematics achievements of TIMSS students is appropriate. Figure 5.2 shows the mean TIMSS mathematics scores associated with students scoring at each point on the 10-point Mathematics Junior Certificate Performance Scale (MJCPS), for all participating students, and students who took the Junior Certificate Examination in 1997. The figure suggests that this scale is an appropriate way to scale the 1996/1997 data also, particularly for grade 8 students (1996 data), where the gradient of the line is quite smooth. There is a mean score difference

between MJCPS scale points of 25.1 scale points on TIMSS mathematics (for all students), or 24.4 scale points (grade 8 students only).

Figure 5.2. Preferred 10-Point MJCPS Plotted Against TIMSS 1995 Mean Mathematics Scores



Note. Scale points 2 and 9 are placeholders, calculated as the midpoint between adjacent means.

In PISA 2000, 2360 out of 3854 students (61.2%) took the examination in 2000; in PISA 2003, 2312 out of 3880 students (59.6%) took the examination in 2003. (Due to missing data on the explanatory variables, 2341 students are in the 2000 dataset, and 2292 students are in the 2003 dataset.) In the case of TIMSS, 2883 out of 6203 students, or 46.5% of the sample took the Junior Certificate in 1996 (due to missing data on the explanatory variables, 2826 students are in the TIMSS dataset). The PISA models include a common set of explanatory variables while the explanatory variables used in the models for TIMSS mathematics and 1996 Junior Certificate mathematics are identical to one another but differ slightly to the PISA models. These are described in more detail in Section 5.6.3.

The procedures used to build the models are similar to those used in the PISA 2000 and PISA 2003 multilevel models for achievement in Ireland (Shiel et al., 2001; Cosgrove et

al., 2005) and are described here in brief. The software HLM 6.0 was used (Raudenbush et al., 2004).

First, all student-level variables were tested separately, and then variables that were significant when tested alone were entered in a model simultaneously. Non-significant variables were removed from the model in sequence using a backwards elimination strategy (i.e., removing non-significant variables, then re-evaluating the model to ensure that all remaining variables retain significance). The school-level variables were then tested in the same way.

Next, all variables at both levels were examined simultaneously, and, if applicable, non-significant variables were removed in sequence. A complicating factor of these models in the case of those involving the PISA and TIMSS achievement data is that each student has five achievement estimates (plausible values, as described in Chapter 1). Fortunately, HLM 6.0 can incorporate plausible values into parameter estimates and variance components. However, the situation is slightly more complex when one is considering the relative improvement in model fit in the case of categorical variables with more than one level (e.g., school sector, which is fitted as two dummy variables). In such instances, one needs to evaluate the change in the deviance (the overall 'fit') of the model with and without the dummy variables, referring this difference to a χ^2 distribution, with degrees of freedom set at the difference in the number of terms in the models compared. Unfortunately, HLM 6.0 does not provide deviance statistics when plausible values are used. Therefore, to evaluate the significance of categorical variables with more than two levels, the models to be compared were computed five times, once with each plausible value, and the averages of the two sets of five deviance statistics compared.

Once the final set of variables was established, the curvilinearity of each continuous variable was explored initially in SPSS 12.0 (e.g., Bryman & Cramer, 2004) by means of ordinary-least-squares regression with the first plausible value as the outcome variable (in the models where plausible values were required), or with performance on the Junior Certificate as the outcome variable (in the models of Junior Certificate performance). In the case of student-level variables, the regression model tested the significance of the original continuous variable plus its squared term (e.g., ESCS +

$ESCS^2$). If the squared term was significant, then it was tested within the model that had been developed using HLM 6.0, comparing the deviance statistic of the model with and without the squared term. In the case of school-level variables, the achievement outcome (either the first plausible value or JCPS score) was aggregated to the school level and ordinary-least-squares regression carried out as per the student-level variables. Any significant squared terms were then evaluated in the same way in HLM 6.0.

When testing for the significance of interactions between variables, it was decided, in the interest of not specifying overly complex models with many terms, to limit tests of interactions to those pertaining to gender at the student level and cross-level interactions involving school-level SES and student gender.

Before finalising the model, the slope of each significant student-level variable was evaluated to see whether its effect was constant across schools (i.e., whether the same student-level model applied across schools) or whether it varied, through the addition of a random coefficient to the slope of each student-level variable tested one at a time.

Some additional points are relevant to the interpretation of the models. Since HLM 6.0 employs listwise deletion, the model parameters pertain only to those students who are not missing data on any variable. Fortunately, missingness is very low. For example, the models of PISA 2003 mathematics exclude just 1.1% of students who are missing data on explanatory variables. Furthermore, the models are unweighted. The explicit sampling stratum is included as a school-level variable because sample weights were not used in the models; inclusion of the sample stratum removes at least some of the variance due to the sample design. Aitkin, Francis and Hinde (2005) argue against the use of weights in model-building for two reasons. First, samples from larger sub-populations are given greater weight, despite the fact that observations are of individuals rather than aggregates. Second, evaluation of the model through examination of the change in deviance when variables are added or removed is affected by the application of weights. This rationale was also used in the multilevel models in the national reports for PISA 2000 and PISA 2003 (e.g., Shiel et al., 2001, p. 97). Thirdly, in the case of the models presented in this chapter, a subset of students is examined and the student weights were computed on the basis of the complete datasets in question, rather than subsets of them, so the appropriateness of using weights would

be questionable. Sample stratum is not used in the TIMSS models since the sample design for Ireland did not use explicit strata (Foy, 1998). Finally, all continuous variables (e.g., student SES) have been centred around their grand mean. This facilitates the interpretation of the intercept since it corresponds to the hypothetical achievement score of a student with an average value on all continuous variables in the model. It also results in greater stability during the estimation process.

I compare six models:

- PISA 2000 reading literacy scores (for students taking the Junior Certificate Examination (JCE) in 2000)
- JCE English (EJCPS 'preferred' scale), 2000
- PISA 2003 mathematics scores for students taking the JCE in 2003
- JCE mathematics (MJCPS 'preferred' scale), 2003
- TIMSS 1995 mathematics scores (for students taking the JCE in 1996)
- JCE mathematics (MJCPS 'preferred' scale), 1996.

5.6.2. Computation of Explained Variance

For each set of models, the percentages of student and school variance explained by each variable when added separately was calculated by school and student economic, social and cultural status (ESCS); by all student-level variables added simultaneously; by all school-level variables added simultaneously; and for the final model. The formula used to calculate the explained variance (Snijders & Bosker, 1999) compares the variance of the null model (i.e., the model with the achievement outcome and no explanatory variables) with that of subsequent models. The formula requires one to use an appropriate value for school enrolment size. As in the national reports for PISA 2000 and PISA 2003, the mean enrolment size of 15-year-olds in schools on the sampling frame was used (86.9 in 2000 and 82.1 in 2003); for TIMSS, the average number of students enrolled in grade 8 in the original list of sampled schools (118.3) was used.

The formula is as follows:

Level 1 $R^2 = 1 - (Var\ L1F + VarL2F)/(VarL1N + VarL2N)$
Level 2 $R^2 = 1 - (Var\ L1F/CS + VarL2F)/(VarL1N/CS + VarL2N)$

Where
VarL1F = Level 1 variance of fitted model

VarL2F = Level 2 variance of fitted model
VarL1N = Level 1 null model variance
VarL2N = Level 2 null model variance
CS = Cluster Size.

5.6.3. Selection and Construction of Explanatory Variables

The variables examined in the models are shown in Table 5.4. While an attempt has been made to render the variables comparable across studies and models, there are some differences which require explanation. The differences occur because the TIMSS database does not include any composite variables and these had to be constructed; the questionnaire items on which they are based differ somewhat to those used in the PISA composites. Also, the choice of variables requires justification.

Student gender is included since there is a known gender difference in achievement with the exception of Junior Certificate mathematics. (Further, gender is included to test for interactions with SES.) The measure of student economic, social and cultural status (ESCS) is a composite which combines parental occupation, parental education, and home possessions relating to wealth and educational resources in the case of models of PISA; the TIMSS measure is similar but excludes parental occupation since these data were not gathered in TIMSS. ESCS is included as a variable of central relevance to the hypotheses under investigation. The inclusion of ESCS as a single composite at the student and school levels is preferable to a host of separate variables, since (i) these may be prone to multicollinearity, which complicates interpretation (Hutcheson & Sofroniou, 1999), and (ii) it is not of interest in the proposed analyses to examine the effects of specific aspects of student social background on achievement.

Table 5.4. Description of Variables Used in the Multilevel Models

School-Level	*Description*
Sampling stratum	Number of 15-year-olds enrolled in the school, used as an explicit stratum to sample schools. Small = up to 40 15-year-olds; Medium = 41-80 15-year-olds; Large = 81+. Entered as two dummy variables (Large and Small stratum indicators). Used in models for 2000 and 2003 only.
School sex compositon	Based on the enrolment of 15-year olds, either single sex (100% boys or 100% girls) or mixed sex. Single sex is the reference group.
School sector	Community/comprehensive, secondary or vocational. Entered as two dummy variables (Community/Comprehensive and Vocational indicators).
School ESCS	Average ESCS scores of the students in the school. Continuous, 1995/1996 M = 0.08, SD = 0.52; 2000 M = -0.11; SD = 0.44; 2003 M = 0.11; SD = 0.46.
School educational resources	Principals' reports on the extent to which student learning is hindered by items such as lack of instructional material, lack of computers, lack of multimedia resources. 1995/1996 M = 0.00, SD = 1.00. 2000 M = 0.16, SD = 1.03. 2003 M = -0.04, SD = 0.85.
Student behaviour	Principals' percentions on how much learning was hindered by factors such as 'student absenteeism', 'disruption of classes by students', 'students lacking respect for teachers'. 2000 M = 0.23, SD = 0.80. 2003 M = -0.26, SD = 0.87.
School autonomy	Principals' reports on which aspects of school management the school had a decision-making role in, such as appointing and dismissing teachers, formulating the school budget, establishing assessment and admittance policies. 1995/1996 M = 0.00, SD = 1.00. 2000 M = 0.01, SD = 0.49. 2003 M = -0.03, SD = 0.47.
Teacher participation	Principals' reports on which aspects of school management the teaching staff had a decision-making role in, such as appointing and dismissing teachers, formulating the school budget, establishing assessment and admittance policies. 1995/1996 M = 0.00, SD = 1.00. 2000 M = 0.41, SD = 0.71. 2003 M = -0.24, SD = 0.68.
School building quality	Principals' reports on the extent to which student learning is hindered by items such as poor condition of buildings and lack of instructional space. 1995/1996 M = 0.00, SD = 1.00. 2000 M = -0.20, SD = 0.93. 2003 M = 0.048 SD = 0.28.
School disciplinary climate	Average of disciplinary climate scores pertaining to the subject of interest - English in 2000 and mathematics in 1995/1996 and 2003. Continuous, 1995/1996 M = 0.08, SD = 0.52; 2000 M = 0.08, SD = 0.38; 2003 M = 0.26, SD = 0.39. Based on students' responses to six Likert-type items such as 'students don't listen to what the teacher says'; 'there is noise and disorder'.
Student-Level	
Gender	Student gender. Female is the reference group.
Student ESCS	Composite combining aspects of students' socioeconomic and social background. These are: higher of parent's occupation, on an international index ranging from 16-90; higher of parental education, according to ISCED classification (ranging from primary to third-level degree - 2000 and 2003 only), and number of home possessions (e.g., mobile phones, cars). Parental occupation not included in the 1995/1996 measure. Continuous, 1995/1996 M = 0.00, SD = 1.00; 2000 M = -0.08; SD = 0.96; 2003M = -0.10; SD = 0.88.

Note. Descriptive statistics are unweighted and based only on the cases in the models.

Three variables relating to school process/climate which were reported by students (disciplinary climate, teacher support in class lessons and student-teacher relations), and eight school-level variables (quality of instructional resources, quality of material resources, perceived shortage of teachers, teacher morale, teacher behaviour, student

behaviour, school autonomy, and teacher participation in decision-making), are available in both the PISA 2000 and PISA 2003 datasets (OECD, 2002b; 2005a). The TIMSS dataset includes student questionnaire items that are comparable to those used in PISA to construct the disciplinary climate measure, but there are no items comparable to those used for the teacher support and student-teacher relations composites. The TIMSS school questionnaire includes some items that can be used to construct comparable measures of five of the eight school-level variables which appear in PISA (school autonomy, teacher participation in decision-making, student behaviour, quality of instructional resources, and quality of material resources). Therefore, each of these was constructed, using principal components analysis as the data reduction method (e.g., Hutcheson & Sofroniou, 1999, pp. 217-251), with the exception of school autonomy and teacher participation, which, similar to the composites for PISA, are based on counts of a number of responses, then re-scaled to have a mean of 0.0 and standard deviation of 1.0. Table A5.1 (Appendix 5) provides a detailed description of the items used to construct the ESCS scale and the six school-level variables for PISA 2000, PISA 2003 and TIMSS 1995. Tables A5.2 to A5.9 provide the factor loadings associated with the items for the TIMSS scales and the counts and transformations for the school autonomy and teacher participation composites. It should be noted that the methods used to construct the composites are not the same in PISA, which used Weighted Likelihood Estimation (Warm, 1989) rather than principal components analysis. Moreover, for all three surveys, all variables collected through the school questionnaire file are missing data. Since HLM 6.0 employs listwise deletion, cases with missing data on these variables are automatically dropped. Therefore, to preserve cases for which data is missing, each variable collected through the school questionnaire has a missing indicator and missing values on the original variable recoded to zero; this is similar to methods used in other explanatory models (e.g., Shiel et al., 2001; Smyth, 1999).

It was suggested in Chapters 1 and 2 that both TIMSS and PISA have been relatively unsuccessful in developing school-level measures relating to school resources, climate, processes, etc. that explain substantial amounts of variance in achievement. A further problem with these measures is that they are difficult to interpret, because they are based on the opinions or perceptions of the principals and students (i.e., constructed from a series of responses to Likert-type items) rather than being objective quantitative

measures, entailing at least some degree of subjective judgement. Also, comparing results across studies places restrictions on the number of variables which may be included. Therefore, the models cannot be regarded as optimal in examining the effects of school resources, climate etc. on achievement.

In addition to school variables relating to processes and climate, two variables relating to structural features (other than sample stratum as noted previously) are also included – sex composition and sector.

Since the construction of the ESCS measure for TIMSS entailed a number of steps it is described here in brief. The TIMSS student questionnaire included questions on parental education, number of books in the home, and possession of a range of educational and material resources. I computed the higher of parents' occupation (which ranges from primary to university degree, available for 86.7% of students) as the first component of the TIMSS ESCS. As the second component, I examined the books in the home measure and noted, similar to the scale used in PISA, that it was not of equal intervals (0-10, 11-25, 26-100, 101-200, >200) so I transformed it to its natural logarithm, which produced a smoother scale. This formed the second component of TIMSS ESCS. As a third component, I carried out a principal components analysis of four educational possessions (calculator, desk, dictionary, and encyclopaedia). These loaded on a single factor (loadings ranging from .44 to .63) and I computed the factor scores based on these four components (mean = 0.0, SD = 1.0). I also examined factor loadings of a range of material possessions, and found that phone, dishwasher, microwave, tumble dryer, and second bathroom loaded on a single factor (loadings from .57 to .73). I computed a factor score for these components also.

Combining these four aspects of ESCS could be done in several ways: I could take the average value, or a weighted average. I explored various ways of combining the four components by examining their correlations with the first plausible value for mathematics and the preferred MJCPS. Table 5.5 shows these correlations. Based on a comparison of these, a composite ESCS which accords twice the weight to parental education and books in the home was selected. To avoid high amounts of missing data on this measure, I computed a weighted average ESCS which excluded the parental

education variable for the students that had missing data on that variable. I then re-scaled the composite to have a mean of 0.0 and a standard deviation of 1.0.

Table 5.5. Pearson Correlations Between Achievement on TIMSS and MJCPS and Four Alternative ESCS Composites

	TIMSS	*MJCPS*
Raw ESCS1	.274	.338
Raw ESCS2	**.304**	**.371**
Raw ESCS3	.299	.364
Raw ESCS4	.291	.364

Note.
Scale 1: average of the four components.
Scale 2: weighted average assigning twice the weight to parental education and books in the home.
Scale 3: weighted average assigning three times the weight to parental education and books in the home.
Scale 4: weighted average assigning twice the weight to home educational resources and material possessions.

5.7. Results

5.7.1. Multilevel Models of Achievement on PISA 2000 Reading

Prior to entering any terms in the model, variance components for the null model were computed to obtain a measure of the total variance in achievement, and the proportion of the total that is between schools, which was found to be 18.4%. The mean reading score of students included in the model is 517.6, and the standard deviation is 91.9.[40] Tables 5.6 and 5.7 show the parameters for the student-level and school-level variables tested separately. The tables show, for example, that females score about 24 points higher than males, on average, and that there is 33-point increase associated with a one standard deviation increase in school ESCS. Student ESCS explains substantial portions of the variance at both student (12.5%) and school (37.0%) levels; school ESCS also explains large portions of the variance (11.7% at the student level and 59.5% at the school level). Three of the school-level variables are not significant (material resources, teachers' participation in decision-making, and school autonomy), and school sample stratum is borderline significant.

Table 5.6. Achievement on PISA 2000 Reading: All Student-Level Variables Tested as Separate Models by Addition to the Null Random Intercept Model

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*	*% student var*	*% sch var*
Gender: female–male	24.300	4.336	t = 5.604	2339	< .001	2.6	10.3
ESCS	27.922	1.654	t = 16.885	2339	<.001	12.5	37.0

[40] These are the means of the five plausible values, unweighted, for the cases included in the model.

Table 5.7. *Achievement on PISA 2000 Reading: All School-Level Variables Tested as Separate Models by Addition to the Null Random Intercept Model*

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*	*% student var*	*% sch var*
Sample Stratum							
Small-Medium	1.369	22.412	Ddiff = 5.522	2	.063	1.0	5.0
Large-Medium	19.966	9.839					
Single Sex-Mixed Sex	28.242	7.056	t = 4.002	136	<.001	2.3	11.8
Sector							
Comm/Comp-Secondary	-26.280	9.563	Ddiff = 54.709	2	<.001	7.6	38.8
Vocational-Secondary	-62.292	8.57					
Average ESCS	72.734	6.309	t = 11.529	136	.001	11.7	59.5
Average Disc. Climate	-33.483	9.473	t = -3.535	136	<.001	1.9	10.1
Material Resources	0.763	3.720	Ddiff = .087	2	.961	0.0	0.1
Missing Mat. Resources	9.429	44.1					
Student Behaviour	-22.11	4.426	Ddiff = 23.186	2	<.001	3.8	19.4
Missing Stud. Behaviour	4.219	40.575					
School Autonomy	31.247	9.59	Ddiff = 16.549	2	<.001	2.5	13.0
Missing Sch. Autonomy	-11.129	15.179					
Teachers' Decision-Making	3.719	5.258	Ddiff = 0.614	2	.736	0.1	0.4
Missing Tch. Decision	10.002	15.317					
School Building Quality	-3.928	4.281	Ddiff = 0.921	2	.631	0.2	0.9
Missing Sch. Building	-2.575	9.958					

Table 5.8 shows the student-level variables entered simultaneously. Both remain highly significant and their parameter estimates are similar to estimates when entered one at a time.

Table 5.8. *Achievement on PISA 2000 Reading: All Student-Level Variables Tested Simultaneously*

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	503.526	3.593	t = 140.138	137	<.001
Gender: Female-Male	26.482	4.199	t = 6.306	2338	<.001
ESCS	28.537	1.630	t = 17.504	2338	<.001

Table 5.9 shows the parameter estimates for all significant school-level variables entered simultaneously. School ESCS and disciplinary climate retain their significance; the other variables are no longer significant. Table 5.10 shows the parameters for school ESCS and disciplinary climate following removal of the non-significant terms.

Table 5.9. *Achievement on PISA 2000 Reading: All School-Level Variables Tested Simultaneously*

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	511.867	6.698	t = 76.423	126	<.001
Sample Stratum					
Small-Medium	-7.308	12.670	Ddiff = 1.170	2	.557
Large-Medium	4.686	5.975			
Single Sex-Mixed Sex					
Sector					
Comm/Comp-Secondary	-0.877	8.571	Ddiff = 1.766	2	.414
Vocational-Secondary	12.938	9.189			
Average ESCS	53.972	7.152	t = 7.547	126	<.001
Average Disc. Climate	-20.153	6.710	t = -3.003	126	.004
Student Behaviour					
Student Behaviour	-7.866	3.197	Ddiff = 4.104	2	.129
Missing Stud. Behaviour	12.512	10.739			
School Autonomy					
School Autonomy	7.674	5.566	Ddiff = 2.368	2	.306
Missing Sch. Autonomy	-7.122	7.309			

Table 5.10. *Achievement on PISA 2000 Reading: All Significant School-Level Variables Tested Simultaneously*

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	515.326	2.597	t = 198.399	135	<.001
Average ESCS	70.731	5.870	t = 12.049	135	<.001
Average Disc. Climate	-26.957	6.333	t = -4.256	135	<.001

All school- and student-level variables were then entered simultaneously. After checking that they retained significance, I tested the curvilinearity of the continuous variables using ordinary-least-squares regression in SPSS. In the case of school disciplinary climate, no evidence of curvilinearity was found; however, both student and school ESCS demonstrated a significant curvilinear trend so they were included as two additional terms. The interaction between gender and student ESCS was not significant, nor was the interaction between gender and school ESCS. As a final check of the model, random components were added to the slopes of the two student-level variables one at a time to see if their effects varied across schools. The effects of both student ESCS and gender were found to be constant. The final model explains 21.4% of within-school variance, and 76.7% of the variance between schools (or 31.6% of the total variance in achievement). The two school-level variables explain an additional 6.0% of within-school variance, and 28.8% of the variance between schools. The final model for PISA 2000 reading is shown in Table 5.11.

Table 5.11. Final Model of Achievement on PISA 2000 Reading

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	503.177	3.297	t = 152.631	134	<.001
Student-Level Variables					
Gender: Female-Male	25.315	3.882	t = 14.537	2334	<.001
ESCS Parameters					
ESCS	25.243	1.737			
ESCS Squared	3.735	1.478	t = 2.527	267	.012
School-Level Variables					
School ESCS Parameters					
Average ESCS	42.327	5.262			
Average ESCS Squared	-30.584	7.786	t = -3.928	134	<.001
Average Disc. Climate	-18.624	6.098	t = -3.054	134	.003
Variance Components					
Intercept variance	307.079				
Level-1 (within-school) variance	6341.522				

The rather weak curvilinear nature of the relationship between student ESCS and achievement is shown in Figure 5.3. It suggests that there is a predicted achievement difference on PISA 2000 reading of about one and one-sixth standard deviations between students with ESCS scores that are two standard deviations above and below the mean.

Figure 5.3. Plot of the Relationship Between Student ESCS and Student Achievement on PISA 2000 Reading

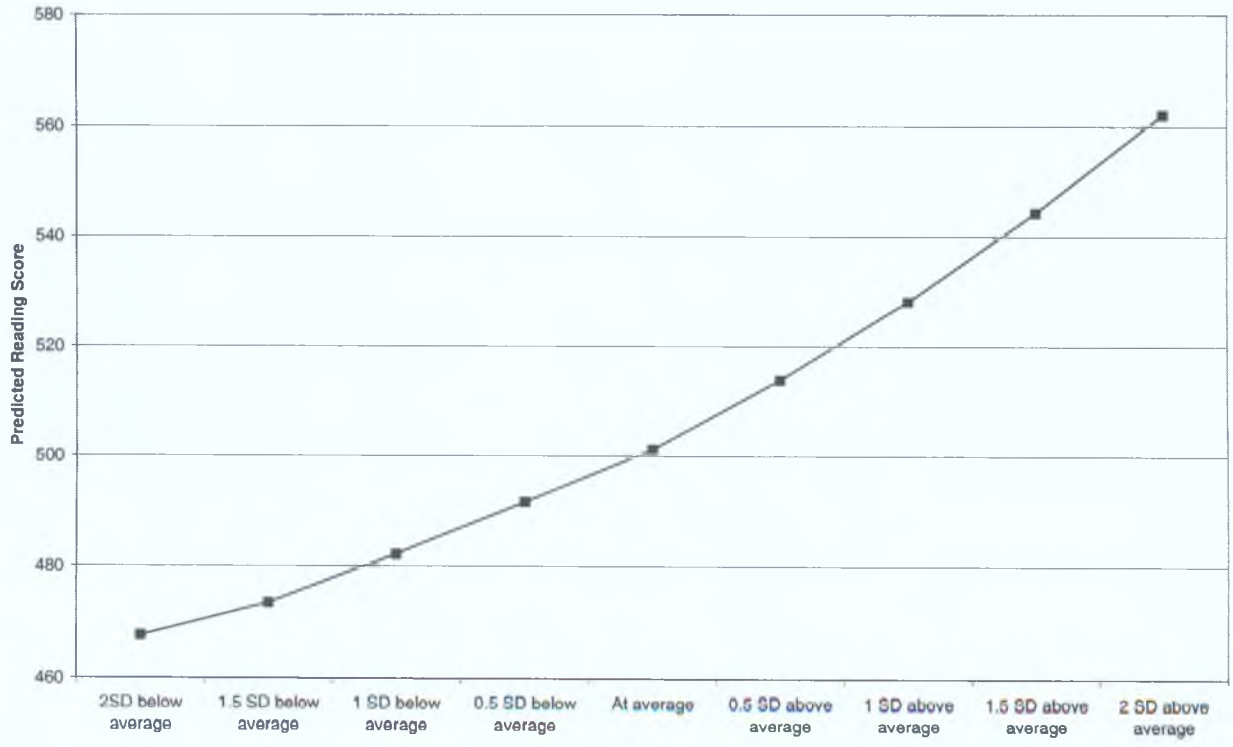
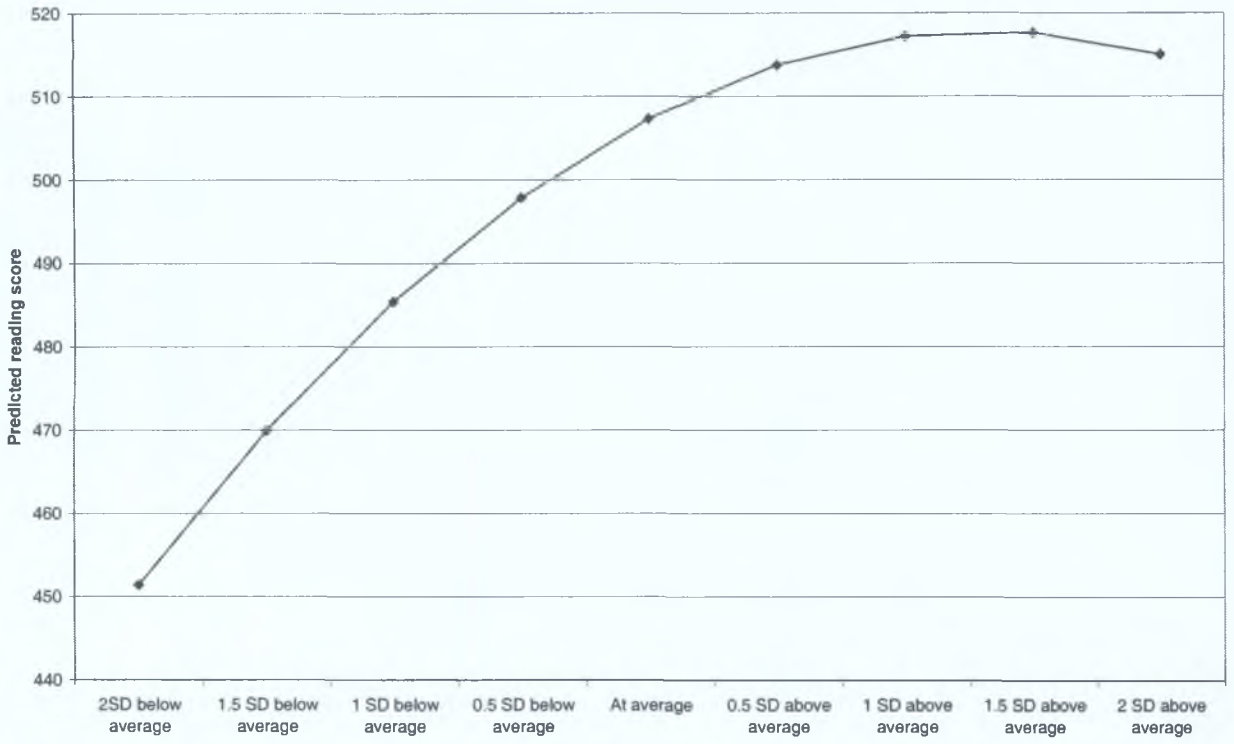


Figure 5.4 shows the curvilinear relationship between school ESCS and achievement, which shows that the social context effect tapers off at higher levels of school mean ESCS.

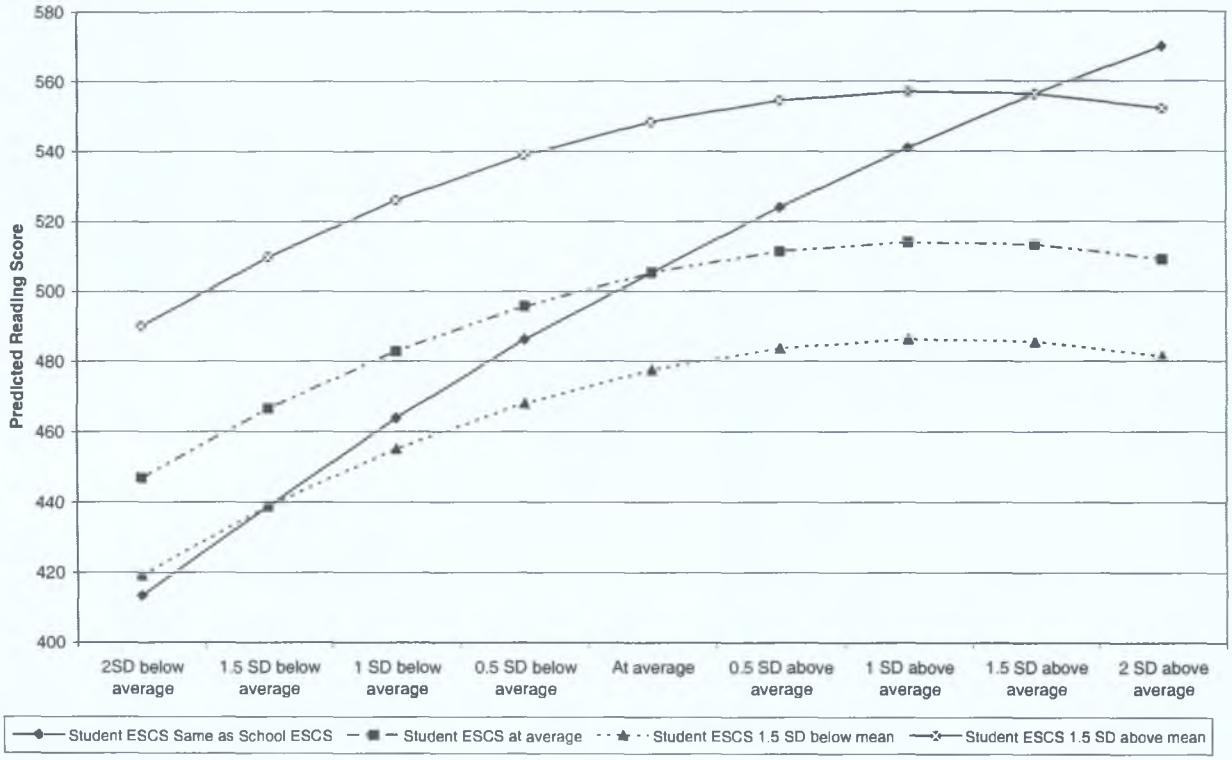
Figure 5.4. Plot of the Relationship Between School ESCS and Student Achievement on PISA 2000 Reading



When one considers the combined effect of school and student ESCS, one can see the detrimental impact relating to individuals of low-ESCS backgrounds in a school of low average ESCS (Figure 5.5). Students whose ESCS score is two standard deviations below the mean and who are in schools with average ESCS scores two standard deviations below the mean have an expected reading score that is about 1.7 standard deviations below that of students of ESCS whose ESCS score is two standard deviations above the mean and in schools with average ESCS scores two standard deviations above the mean. The figure also suggests that students with ESCS scores 1.5 standard deviations above the average in a school which has a mean ESCS 1.5 standard deviations below the average have an expected reading score that is about the same as a student with an average ESCS score in a school with an ESCS score 1.5 standard deviations above the mean.

In Figures 5.3 to 5.5 (and for all subsequent plots of ESCS), the expected scores of females are plotted. However, there is no gender interaction present in the model, so the expected score differences hold equally for males. The parameter estimate for the gender variable indicates that the predicted scores for males along any of the points plotted are about 25 scale points lower than those for females.

Figure 5.5. Plot of the Relationship Between the Combined Effect of Student and School ESCS and Student Achievement on PISA 2000 Reading



5.7.2. Multilevel Models of Achievement on Junior Certificate English for Students Participating in PISA 2000

Prior to entering any terms in the model, variance components associated with the null model were computed to obtain a measure of the total variance that is between schools, which was found to be 20.6%. The mean for the EJCPS is 6.04 and the standard deviation 1.76. Tables 5.12 and 5.13 show the parameters for the student-level and school-level variables tested separately. The gender difference, favouring females, is significant; the effects for both student and school ESCS are again substantial. At the school level, stratum, sex composition, sector, ESCS, disciplinary climate, and student behaviour are significant.

Table 5.12. *Achievement on ECJPS, 2000: All Student-Level Variables Tested as Separate Models by Addition to the Null Random Intercept Model*

	Parameter	SE	Test Statistic	df	p–value	% student var	% sch var
Gender: female–male	0.762	0.098	t = 7.768	2339	<.001	5.6	17
ESCS	0.554	0.033	t = 16.593	2339	<.001	13.7	36.5

Table 5.13. *Achievement on EJCPS, 2000: All School-Level Variables Tested as Separate Models by Addition to the Null Random Intercept Model*

	Parameter	SE	Test Statistic	df	p–value	% student var	% sch var
Sample Stratum							
Small-Medium	0.149	0.360	Ddiff = 8.395	2	.015	1.4	6.7
Large-Medium	0.496	0.212					
Single Sex-Mixed Sex	0.669	0.137	t = 4.884	136	<.001	3.5	16.5
Sector							
Comm/Comp-Secondary	-0.554	0.165	Ddiff = 54.814	2	<.001	8.2	38.1
Vocational-Secondary	-1.233	0.156					
Average ESCS	1.404	0.127	t = 11.034	136	<.001	11.9	54.9
Average Disc. Climate	-0.634	0.202	t = -3.160	136	.002	1.9	9.0
Material Resources	-0.087	0.073	Ddiff = 1.966	2	.374	0.4	1.6
Missing Mat. Resources	0.655	0.872					
Student Behaviour	-0.445	0.087	Ddiff = 24.511	2	<.001	4.1	18.9
Missing Stud. Behaviour	0.566	0.803					
School Autonomy	0.387	0.188	Ddiff = 6.307	2	.043	1.2	5.7
Missing Sch. Autonomy	0.247	0.313					
Teachers' Decision-Making	0.069	0.111	Ddiff = 0.563	2	.755	0.1	0.5
Missing Tch. Decision	0.27	0.313					
School Building Quality	-0.155	0.091	Ddiff = 3.774	2	.152	0.7	3.2
Missing Sch. Building	0.261	0.327					

Table 5.14 shows the student-level variables entered simultaneously. Both remain significant and their parameter estimates are similar to when entered one at a time.

Table 5.14. *Achievement on EJCPS, 2000: All Student-Level Variables Tested Simultaneously*

	Parameter	SE	Test Statistic	df	p–value
Intercept	5.643	0.735	t = 76.787	137	<.001
Gender: Female-Male	0.798	0.090	t = 8.909	2338	<.001
ESCS	0.572	0.033	t = 17.590	2338	<.001

Table 5.15 shows the parameter estimates for all significant school-level variables entered simultaneously. School sex composition, ESCS and disciplinary climate retain their significance; the other variables are no longer significant. The direction of the parameter estimate for sex composition suggests that students in single-sex schools do significantly better than those in mixed-sex schools. Table 5.16 shows the parameters

for school ESCS, disciplinary climate and school sex composition following removal of the non-significant terms.

Table 5.15. Achievement on EJCPS, 2000: All School-Level Variables Tested Simultaneously

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	5.831	0.133	t = 43.732	126	<.001
Sample Stratum					
Small-Medium	0.073	0.253	Ddiff = 1.839	2	.399
Large-Medium	0.166	0.122			
Single Sex-Mixed Sex	0.316	0.122	t = 2.590	126	.011
Sector					
Comm/Comp-Secondary	-0.057	0.170	Ddiff = 3.360	2	.186
Vocational-Secondary	-0.308	0.175			
Average ESCS	1.035	0.145	t = 7.137	126	<.001
Average Disc. Climate	-0.421	0.141	t = -2.975	126	.004
Student Behaviour					
Student Behaviour	-0.15	0.071	Ddiff = 4.464	2	.107
Missing Stud. Behaviour	-0.017	0.802			
School Autonomy					
School Autonomy	-0.105	0.114	Ddiff = 1.352	2	.509
Missing Sch. Autonomy	0.443	0.578			

Table 5.16. Achievement on EJCPS, 2000: All Significant School-Level Variables Tested Simultaneously

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	5.844	0.067	t = 87.627	134	<.001
Single Sex-Mixed Sex	0.406	0.105	t = 3.861	134	<.001
Average ESCS	1.247	0.118	t = 10.528	134	<.001
Average Disc. Climate	-0.538	0.128	t = -4.216	134	<.001

There is no evidence of curvilinearity in the continuous explanatory variables student ESCS and school disciplinary climate. However, school ESCS shows a significant curvilinear trend. The interaction between gender and student ESCS is not significant, nor is the interaction between gender and school ESCS. When a test of the constancy of slope variation across schools for the student variables was made by introducing an error term to the slope for gender and ESCS one at a time in the model, the results indicate that the slopes are constant across schools for ESCS, but that the slope for gender varies significantly across schools.

I investigated whether I could model the variation in the slope for gender using each of the school-level variables in turn and none is significant with the exception of school building quality, which, given that higher values on this composite represent poorer

building quality, suggests that the gender difference may be smaller in schools where the quality of the school building is poor. Table 5.17 shows a model for achievement on EJCPS which excludes school building quality as an explanatory variable for the slope for gender, and Table 5.18 shows the final model with these terms included. School sex composition changes from borderline significant to significant with the addition of school building quality to the slope for gender.

To obtain the range of values associated with student gender in approximately 95% of the schools, one can take the square root of the variance of the random slope and add $\pm$ 1.96 times this to the parameter estimate. The square root of the variance is 0.467. The likely range of values associated with the gender difference is therefore -0.159 to 1.672. The final model explains 24.8% of within-school variance, and 77.2% of the variance between schools (or 36.6% of the total variance in achievement). The school-level variables explain an additional 5.8% of within-school variance, and 24.6% of the variance between schools.

Table 5.17. Model of Achievement on EJCPS 2000, Without Explanatory Variable for Student Gender Slope Variation

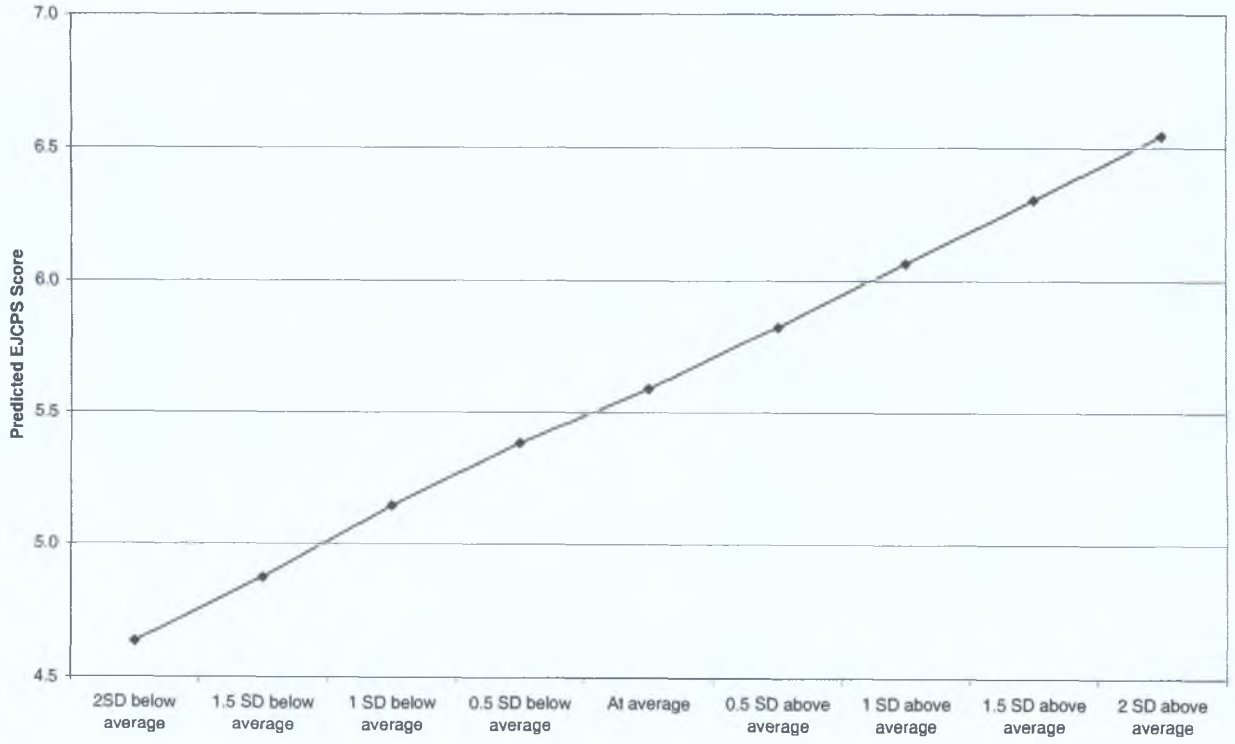
	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	5.579	0.069	t = 80.575	133	<.001
Student-Level Variables					
Gender: Female-Male	0.756	0.083	t = 9.101	137	<.001
ESCS	0.501	0.035	t = 14.130	2334	<.001
School-Level Variables					
Single Sex-Mixed Sex	0.170	0.093	t = 1.818	133	.071
ESCS Parameters					
Average ESCS	0.741	0.117			
Average ESCS Squared	-0.508	0.184	t = -2.760	133	.007
Average Disc. Climate	-0.302	0.118	t = -2.555	133	.012
Variance Components					
Intercept variance	0.181				
Gender slope variance	0.226				
Level-1 (within-school) variance	2.172				

Table 5.18. Final Model of Achievement on EJCPS 2000

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	5.572	0.068	t = 81.388	133	<.001
Student-Level Variables					
Gender: Female-Male	0.759	0.082	t = 9.207	135	<.001
Gender slope variance					
School Building Quality	-0.169	0.054	Ddiff = 8.516	2	.014
Missing Sch. Building	0.006	0.088			
ESCS	0.499	0.035	t = 14.084	2332	<.001
School-Level Variables					
ESCS Parameters					
Average ESCS	0.720	0.109			
Average ESCS Squared	-0.473	0.167	t = -2.839	133	.006
Average Disc. Climate					
Variance Components					
Intercept variance	0.180				
Gender slope variance	0.218				
Level-1 (within-school) variance	2.171				

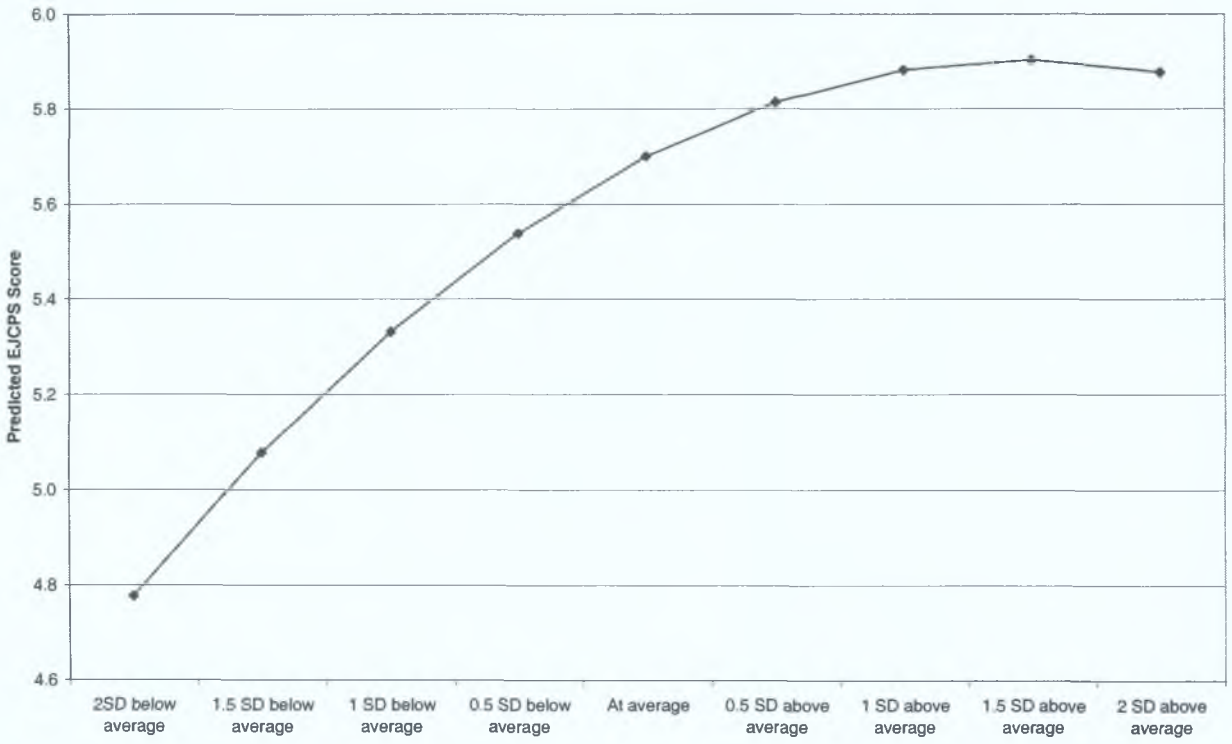
The relationship between student ESCS and achievement is shown in Figure 5.6. It indicates that about 1.2 standard deviations on the EJCPS separate students of ESCS two standard deviations above and below the mean.

Figure 5.6. Plot of the Relationship Between Student ESCS and Student Achievement on EJCPS, 2000



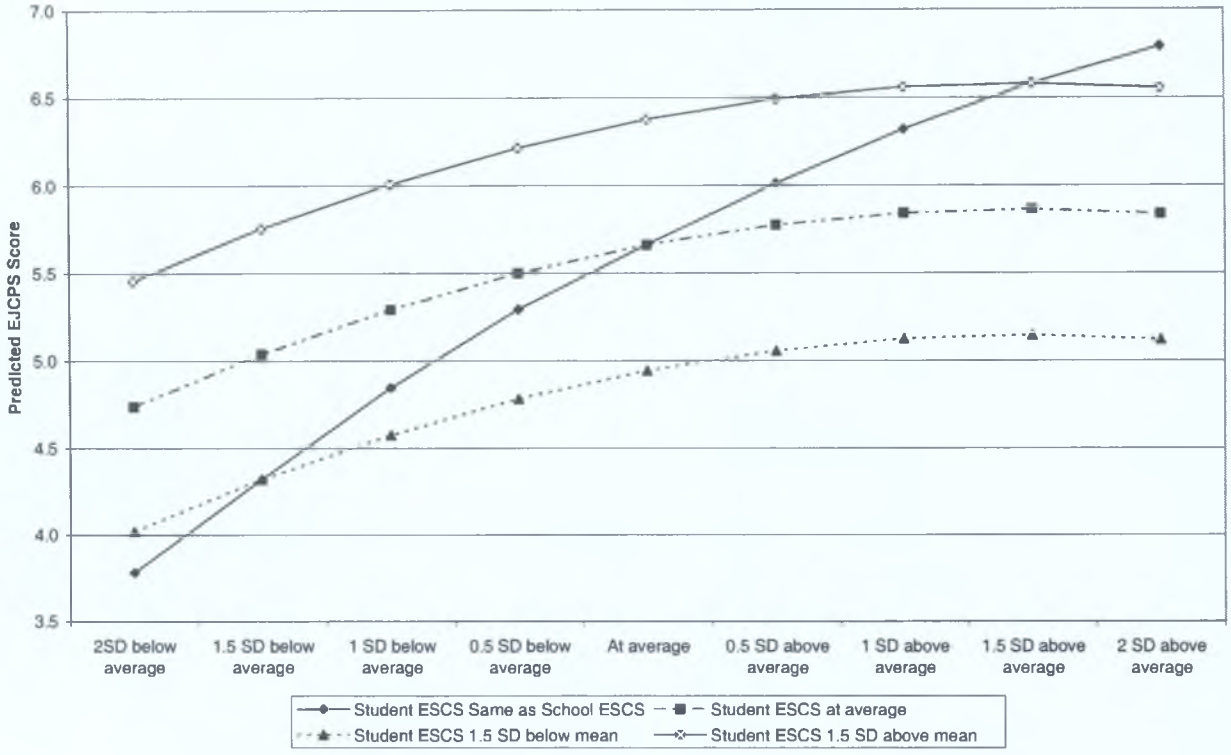
The curvilinear nature of the relationship between school ESCS and achievement is shown in Figure 5.7. It indicates that there is about three-fifths of a standard deviation on the predicted EJCPS scores of students in schools with a mean ESCS two standard deviations above and below the mean. The figure also shows that the school context effect is much weaker at the upper end of the school ESCS distribution.

Figure 5.7. Plot of the Relationship Between School Mean ESCS and Student Achievement on EJCPS, 2000



The combined school and student effects are shown in Figure 5.8, which shows that the predicted difference in EJCPS scores of students with an ESCS score that is two standard deviations below the average, and in schools with a mean ESCS that is two standard deviations below the average, compared to students with an ESCS score two standard deviations above the average and in schools with a mean ESCS two standard deviations above the average is about 1.7 standard deviations. Note that these plots are for females; plots for males take the same form but each plotted point is 0.759 EJCPS points lower.

Figure 5.8. Plot of the Relationship Between the Combined Effect of Student and School ESCS and Student Achievement on EJCPS, 2000



5.7.3. Multilevel Models of Achievement on PISA 2003 Mathematics

Prior to entering any terms in the model, variance components of the null model were computed to obtain a measure of the total variance that is between schools (14.8%). The mean mathematics score of students included in the model is 493.2, and the standard deviation is 82.1.[41] Tables 5.19 and 5.20 show the parameters for the student-level and school-level variables tested separately. The tables show, for example, that females score about 16 points lower than males, on average, and that there is an estimated 67-point increase associated with a one standard deviation increase in school ESCS. All variables with the exceptions of school disciplinary climate, material resources, and teachers' decision-making, are significant. Student ESCS explains a substantial portion of the variance at student (15.4%) and school (53.9%) levels; school ESCS also explains large portions of the variance (12.0% at the student level and 75.1% at the school level).

Table 5.19. Achievement on PISA 2003 Mathematics: All Student-Level Variables Tested as Separate Models by Addition to the Null Random Intercept Model

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*	*% student var*	*% sch var*
Gender: female–male	-15.989	4.425	t = -3.614	2290	.001	0.7	0.6
ESCS	32.291	1.952	t = 16.541	250	<.001	15.4	53.9

[41] These are the means of the five plausible values, unweighted.

Table 5.20. *Achievement on PISA 2003 Mathematics: All School-Level Variables Tested as Separate Models by Addition to the Null Random Intercept Model*

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*	*% student var*	*% sch var*
Sample Stratum							
Small-Medium	-15.736	21.877	Ddiff = 11.346	2	<.001	1.5	8.9
Large-Medium	19.104	7.419					
Single Sex-Mixed Sex	26.037	6.015	t = 4.328	140	<.001	2.5	15.6
Sector							
Comm/Comp-Secondary	-13.450	6.641	Ddiff = 27.353	2	<.001	3.5	21.8
Vocational-Secondary	-37.813	7.799					
Average ESCS	66.668	5.633	t = 11.835	140	<.001	12.0	75.1
Average Disc. Climate	13.603	8.489	t = 1.602	140	.111	0.4	2.4
Material Resources	-3.595	3.556	Ddiff = 1.705	2	.426	0.3	1.8
Missing Mat. Resources	-8.813	11.987					
Student Behaviour	11.070	3.607	Ddiff = 10.616	2	.005	1.4	8.9
Missing Stud. Behaviour	-15.715	11.502					
School Autonomy	21.983	6.476	Ddiff = 12.004	2	.003	1.7	10.5
Missing Sch. Autonomy	-13.395	10.874					
Teachers' Decision-Making	-3.242	4.423	Ddiff = 2.002	2	.368	0.3	2.1
Missing Tch. Decision	-13.507	11.531					
School Building Quality	-8.080	3.272	Ddiff = 7.970	2	.019	1.1	7.2
Missing Sch. Building	-7.136	12.074					

Table 5.21 shows the student-level variables entered simultaneously. Both remain highly significant and their parameter estimates are similar to when entered one at a time.

Table 5.21. *Achievement on PISA 2003 Mathematics: All Student-Level Variables Tested Simultaneously*

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	501.560	3.070	t = 163.401	141	<.001
Gender: Female-Male	-16.762	4.051	t = -4.138	2289	<.001
ESCS	32.512	1.971	t = 16.493	282	<.001

Table 5.22 shows the parameter estimates for all significant school-level variables entered simultaneously. School ESCS and school building quality retain significance; the remainder of the school-level variables are no longer significant. Table 5.23 shows the parameters for school ESCS and school building quality following removal of the non-significant terms.

Table 5.22. *Achievement on PISA 2003 Mathematics: All School-Level Variables Tested Simultaneously*

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	486.565	6.157	t = 79.030	130	<.001
Sample Stratum					
Small-Medium	-5.330	9.487	Ddiff = 1.071	2	.585
Large-Medium	4.163	5.223			
Single Sex-Mixed Sex	5.695	5.325	t = 1.070	130	.287
Sector					
Comm/Comp-Secondary	7.276	6.03	Ddiff = 2.832	2	.243
Vocational-Secondary	-0.423	6.291			
Average ESCS	62.763	6.316	t = 9.937	130	<.001
Student Behaviour					
Student Behaviour	2.822	2.646	Ddiff = 0.994	2	.608
Missing Stud. Behaviour	-29.051	7.984			
School Autonomy					
School Autonomy	-7.278	4.615	Ddiff = 2.993	2	.224
Missing Sch. Autonomy	-29.051	7.984			
School Building Quality					
School Building Quality	-5.860	2.094	Ddiff 9.844	2	.007
Missing Sch. Building	25.454	11.058			

Note. The same schools were missing both student behaviour and school autonomy.

Table 5.23. *Achievement on PISA 2003 Mathematics: All Significant School-Level Variables Tested Simultaneously*

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	492.990	2.067	t = 238.490	138	<.001
Average ESCS	65.247	5.146	t = 12.680	138	<.001
School Building Quality					
School Building Quality	-4.874	2.047	Ddiff = 6.767	2	.034
Missing Sch. Building	-3.883	9.009			

All school- and student-level variables were then entered simultaneously. After checking that they retained significance, I tested the curvilinearity of the continuous variables. In the case of student ESCS and school building quality, no evidence of curvilinearity was found; however, school ESCS demonstrated a significant curvilinear trend so the square of ESCS was included as an additional term. Upon addition of this term, the effect associated with school building quality is only borderline significant (Deviance difference = 4.658, df = 2, p = .097) so this variable was removed from the model. The interaction between student gender and student ESCS was not significant, nor was the interaction between student gender and school ESCS. Random components were added to the slopes of the two student-level variables one at a time to see if their effects varied across schools. The effect of student ESCS was found to be constant, while the slope associated with student gender was varied significantly across schools. I investigated whether I could model the variance in the slope for gender using each of

the school-level variables in turn, but none was significant. Other factors not accounted for in the model are responsible for the between-school variance in the gender slope. Taking the square root of the variance of the random slope and adding $\pm$ 1.96 to the parameter estimate gives the range of values associated with student gender in about 95% of the schools. The square root of the variance is 22.9. The likely range of values associated with the gender difference is therefore -59.7 to 31.0. The final model explains 20.0% of within-school variance, and 77.2% of the variance between schools (or 28.5% of the total variance in achievement). The school-level variable explains an additional 22.2% of between-school variance, and 3.8% of the variance within schools. The final model is shown in Table 5.24.

Table 5.24. Final Model of Achievement on PISA 2003 Mathematics

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	499.473	2.836	t = 176.130	139	<.001
Student-Level Variables					
Gender: Female-Male	-14.757	3.809	t = 3.875	141	<.001
ESCS	27.972	2.058	t = 13.591	567	<.001
School-Level Variables					
ESCS Parameters					
Average ESCS	35.473	5.473			
Average ESCS Squared	-21.318	2.756	t = -3.704	139	.001
Variance Components					
Intercept variance	345.177				
Gender slope variance	524.111				
Level-1 (within-school) variance	5171.684				

Figure 5.9 plots the relationship between student ESCS and achievement on PISA 2003 mathematics. It indicates that the expected score difference between a student with an ESCS score two standard deviations below the mean and a student with an ESCS two standard deviations above the mean is about 100 score points, or 1.2 standard deviations, on the PISA 2003 mathematics scale.

The curvilinear nature of the relationship between school ESCS and achievement is shown in Figure 5.10. The figure indicates that the effect of school ESCS on achievement is much weaker in schools with high average ESCS compared to those with low average ESCS. Overall, about two-thirds of a standard deviation separates the achievement scores of students in schools with ESCS two standard deviations above and below the mean.

Figure 5.9. Plot of the Relationship Between Student ESCS and Student Achievement on PISA 2003 Mathematics

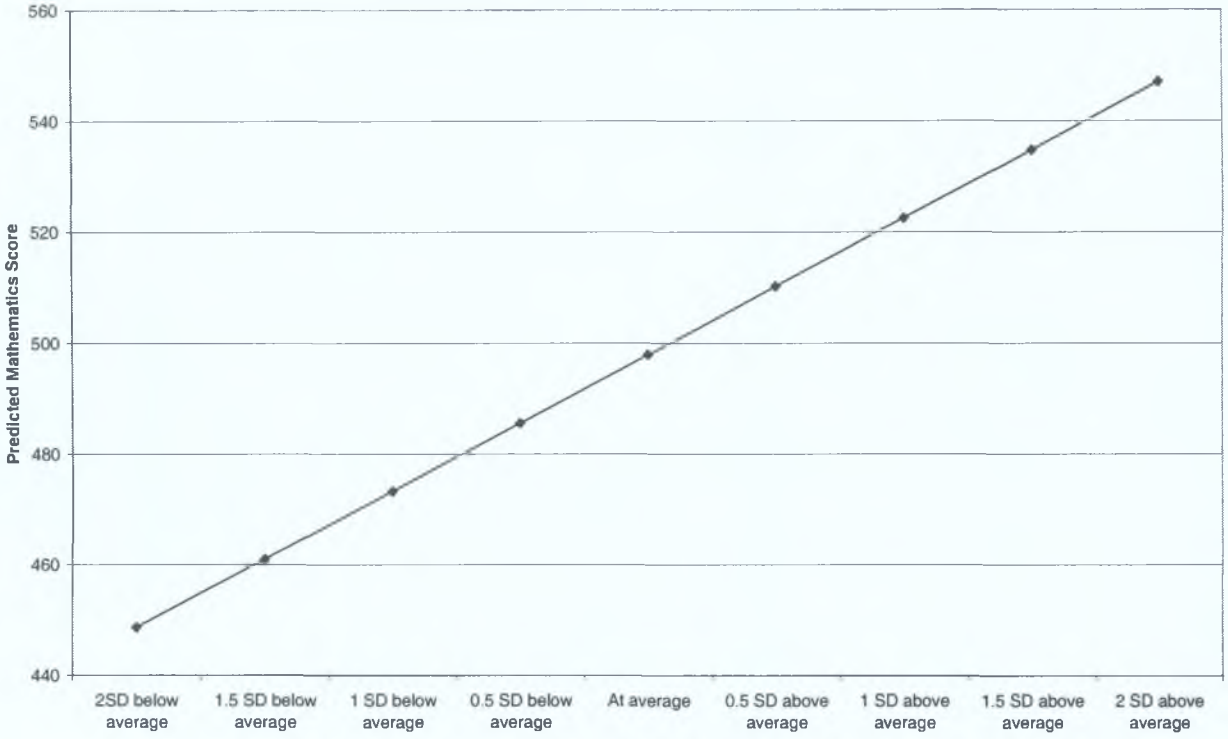


Figure 5.10. Plot of the Relationship Between School Mean ESCS and Student Achievement on PISA 2003 Mathematics

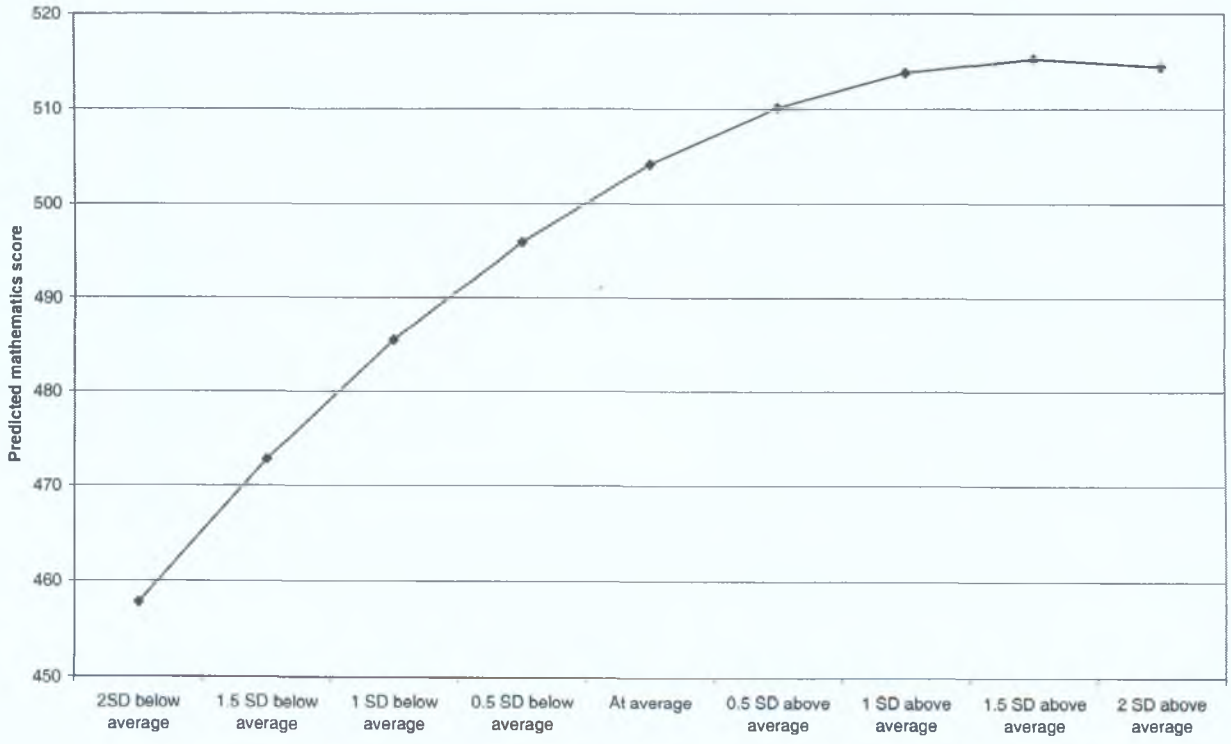
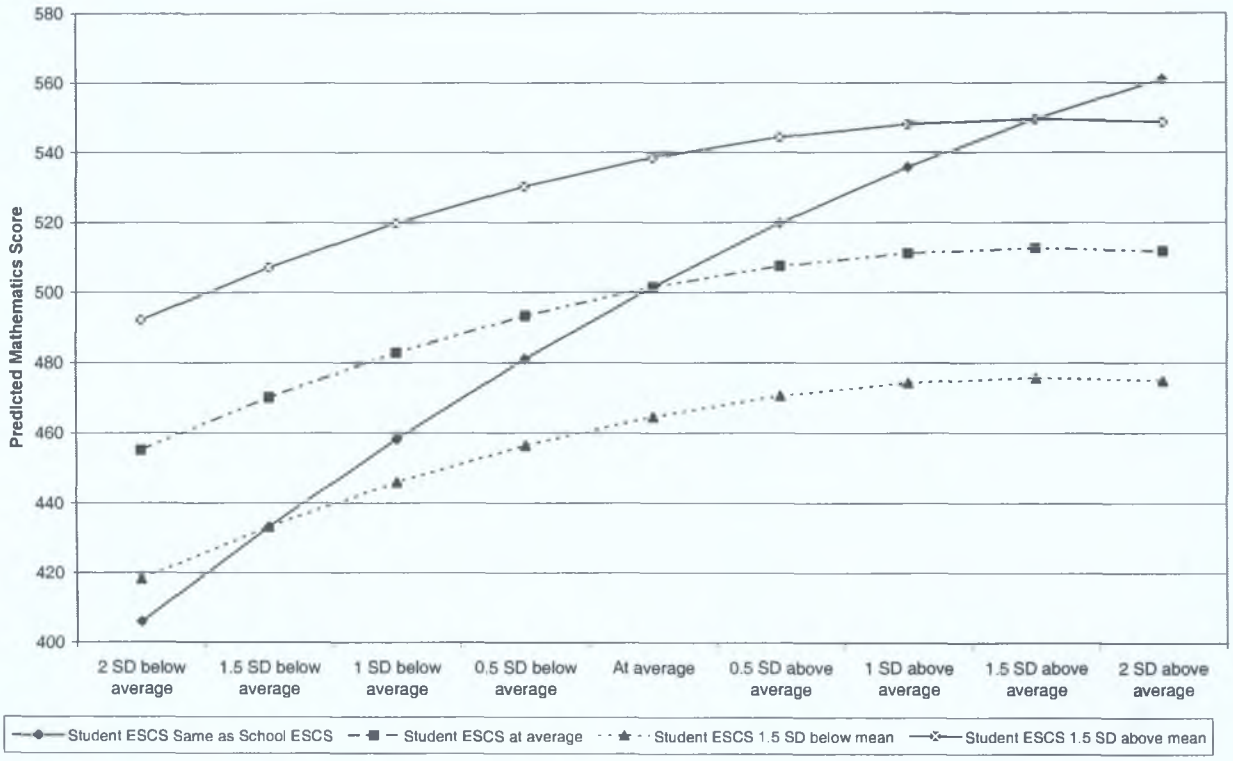


Figure 5.11 considers the combined effects of student and school ESCS. This further illustrates the relative disadvantage of students of low ESCS in schools with low average ESCS, where close to two standard deviations on the PISA 2003 mathematics scale separates students with an ESCS score that is two standard deviations below the mean and who are in schools with average ESCS that is two standard deviations below the mean from students with an ESCS score two standard deviations above the mean and who are in schools with average ESCS two standard deviations above the mean.

Figure 5.11. Plot of the Relationship Between the Combined Effect of Student and School ESCS and Student Achievement on PISA 2003 Mathematics



5.7.4. Multilevel Models of Achievement on Junior Certificate Mathematics for Students Participating in PISA 2003

Prior to entering any terms in the model, variance components associated with the null model were computed to obtain a measure of the total variance that is between schools (20.2%). The mean of the MJCPS for students included in the model is 5.31 and the standard deviation is 2.38. Tables 5.25 and 5.26 show the parameters for the student-level and school-level variables tested separately. There is no significant gender difference; the effect for student ESCS is substantial, with an increase of almost one score point on the MJCPS (or 0.4 of a standard deviation) for a one standard deviation increase on the ESCS scale. All school-level variables with the exceptions of material

resources and teacher participation in decision-making are significant, and the effect associated with school ESCS is again substantial, whereby it explains about three-quarters of the achievement variance between schools.

Table 5.25. *Achievement on MJCPS, 2003: All Student-Level Variables Tested as Separate Models by Addition to the Null Random Intercept Model*

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*	*% student var*	*% sch var*
Gender: female–male	0.128	0.127	t = 1.004	2290	.316	0.2	0.8
ESCS	0.971	0.053	t = 18.177	2290	<.001	18.5	52.7

Table 5.26. *Achievement on MJCPS, 2003: All School-Level Variables Tested as Separate Models by Addition to the Null Random Intercept Model*

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*	*% student var*	*% sch var*
Sample Stratum							
Small-Medium	-0.492	0.581	Ddiff = 11.627	2	<.001	1.9	8.7
Large-Medium	0.636	0.243					
Single Sex-Mixed Sex	0.959	0.187	t = 5.130	140	<.001	3.9	18.4
Sector							
Comm/Comp-Secondary	-0.477	0.196	Ddiff = 35.594	2	<.001	5.6	26.3
Vocational-Secondary	-1.385	0.250					
Average ESCS	2.220	0.163	t = 13.566	140	<.001	16.1	75.3
Average Disc. Climate	0.859	0.268	t = 3.205	140	.002	2.0	9.3
Material Resources	-0.040	0.122	Ddiff = 1.221	2	.543	0.3	1.3
Missing Mat. Resources	-0.385	0.347					
Student Behaviour	0.443	0.108	Ddiff = 16.278	2	<.001	2.7	12.4
Missing Stud. Behaviour	-0.578	0.323					
School Autonomy	0.895	0.203	Ddiff = 19.812	2	<.001	3.3	15.4
Missing Sch. Autonomy	-0.485	0.323					
Teachers' Decision-Making	-0.044	0.145	Ddiff = 1.731	2	.421	0.4	1.8
Missing Tch. Decision	-0.465	0.327					
School Building Quality	-0.227	0.105	Ddiff = 6.605	2	.034	1.2	5.6
Missing Sch. Building	-0.335	0.349					

Table 5.27 shows the student-level variables entered simultaneously. ESCS remains highly significant and gender is not significant in the presence of ESCS; their parameter estimates are similar to when entered one at a time. (Gender is retained for the moment in order to test for interactions.)

Table 5.27. *Achievement on MJCPS, 2003: All Student-Level Variables Tested Simultaneously*

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	5.258	0.900	t = 58.394	141	<.001
Gender: Female-Male	0.100	0.116	t = 0.869	2289	.385
ESCS	0.971	0.153	t = 18.175	2289	<.001

Table 5.28 shows the parameter estimates for all school-level variables entered simultaneously. School stratum, sex composition, school autonomy and student behaviour are no longer significant. School ESCS and disciplinary climate retain their significance, as do sector and school building quality. The direction of the parameter estimates indicate that, after adjusting for the other variables in the model, students in community/comprehensive schools significantly outperform students in secondary and vocational schools. Table 5.29 shows the parameters for the significant school-level variables following removal of the non-significant terms.

Table 5.28. Achievement on MJCPS, 2003: All School-Level Variables Tested Simultaneously

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	5.039	0.172	t = 29.324	129	<.001
Sample Stratum					
Small-Medium	0.157	0.306	Ddiff = 1.213	2	.545
Large-Medium	0.139	0.130			
Single Sex-Mixed Sex	0.176	0.160	t = 1.100	129	.274
Sector					
Comm/Comp-Secondary	0.423	0.187	Ddiff = 7.990	2	.018
Vocational-Secondary	-0.026	0.195			
Average ESCS	2.027	0.160	t = 12.632	129	<.001
Average Disc. Climate	0.755	0.145	t = 5.220	129	<.001
Student Behaviour					
Student Behaviour	0.099	0.068	Ddiff = 2.085	2	.353
Missing Stud. Behaviour	-0.267	0.652			
School Autonomy					
School Autonomy	-0.090	0.134	Ddiff = 0.450	2	.799
Missing Sch. Autonomy	-0.267	0.652			
School Building Quality					
School Building Quality	-0.161	0.055	Ddiff = 8.761	2	.013
Missing Sch. Building	0.027	0.673			

Table 5.29. Achievement on MJCPS, 2003: All Significant School-Level Variables Tested Simultaneously

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	5.266	0.077	t = 68.597	135	<.001
Average ESCS	2.103	0.173	t = 12.142	135	<.001
Average Disc. Climate	0.771	0.128	t = 6.036	135	<.001
Sector					
Comm/Comp-Secondary	0.331	0.152	Ddiff = 8.090	2	.018
Vocational-Secondary	-0.141	0.154			
School Building Quality					
School Building Quality	-0.141	0.056	Ddiff = 8.684	2	.013
Missing Sch. Building	-0.240	0.245			

The curvilinearity of continuous explanatory variables was then tested. Similar to the model for PISA 2003 mathematics, there is no evidence of curvilinearity for student ESCS, school building quality or school disciplinary climate. However, school ESCS shows a significant curvilinear trend. The interaction between gender and student SES is

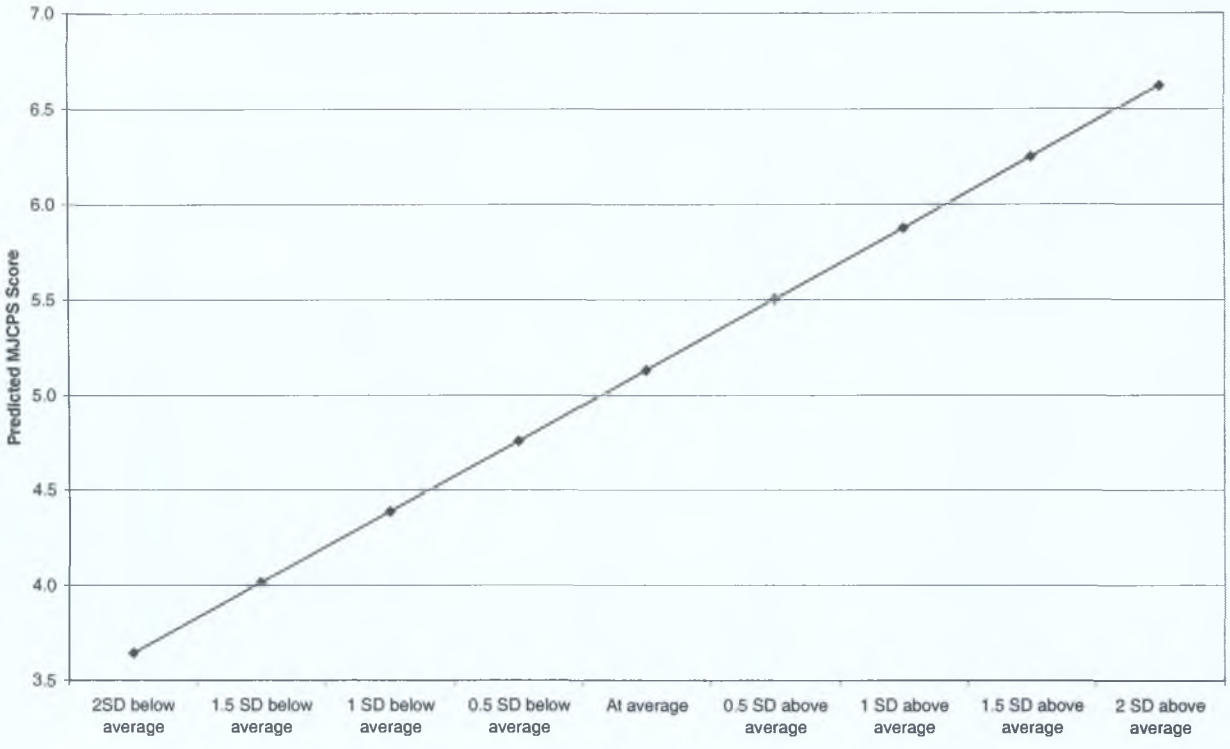
not significant. A test of the constancy of slope variance across schools for the student variables was made by introducing an error term to the slope for gender and ESCS one at a time in the model, and the results indicate that the slopes are constant across schools for both ESCS and gender. The existence of a cross-level interaction between student gender and the school-level ESCS was tested and none was found. Gender was then removed from the model, since it is not itself significant; nor does it contribute to any interactions. Table 5.30 shows the final model of achievement on MJCPS 2003. The model explains 26.0% of within-school variance, and 86.7% of the variance between schools (or 38.3% of the total variance in achievement). The school-level variables explain an additional 7.5% of within-school variance, and 34.0% of the variance between schools.

Table 5.30. Final Model of Achievement on MJCPS 2003

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	5.254	0.074	t = 70.689	134	<.001
Student-Level Variables					
ESCS intercept	0.847	0.055	t = 15.422	2283	<.001
School-Level Variables					
Sector					
Comm/Comp-Secondary	0.307	0.144	Ddiff = 6.002	2	.050
Vocational-Secondary	-0.065	0.145			
ESCS Parameters					
Average ESCS	1.218	0.163			
Average ESCS Squared	-0.591	0.210	t = -2.808	134	.006
Average Disc. Climate	0.783	0.114	t = 6.895	134	<.001
School Building Quality					
School Building Quality	-0.120	0.054	Ddiff = 6.558	2	.038
Missing Sch. Building	-0.190	0.253			
Variance Components					
Intercept variance	0.111				
Level-1 (within-school) variance	4.139				

The relationship between student ESCS and achievement on MJCPS 2003 is shown in Figure 5.12. There is a difference on the MJCPS scale of three scale points, or 1.25 standard deviations, between students with an ESCS score two standard deviations below the mean and students with an ESCS score two standard deviations above the mean.

Figure 5.12. Plot of the Relationship Between Student ESCS and Student Achievement on MCJPS, 2003



The curvilinear nature of the relationship between school ESCS and achievement is shown in Figure 5.13. It shows how the increase in achievement associated with attending a higher ESCS school tapers off at around one standard deviation above the school ESCS mean. The expected MJCPS score difference between the lowest and highest ESCS points is around five-sixths of a standard deviation (or about two MJCPS scale points).

The combined school and student effects are shown in Figure 5.14, which indicates that the predicted difference in MJCPS scores of students with ESCS scores two standard deviations below the average, and who are in schools with a mean ESCS that is two standard deviations compared to students with ESCS scores that are two standard deviations above the average, and who are in schools with a mean ESCS that is two standard deviations above the average is around 5 MJCPS scale points, or just over two standard deviations.

Figure 5.13. Plot of the Relationship Between School Mean ESCS and Student Achievement on MJCPS, 2003

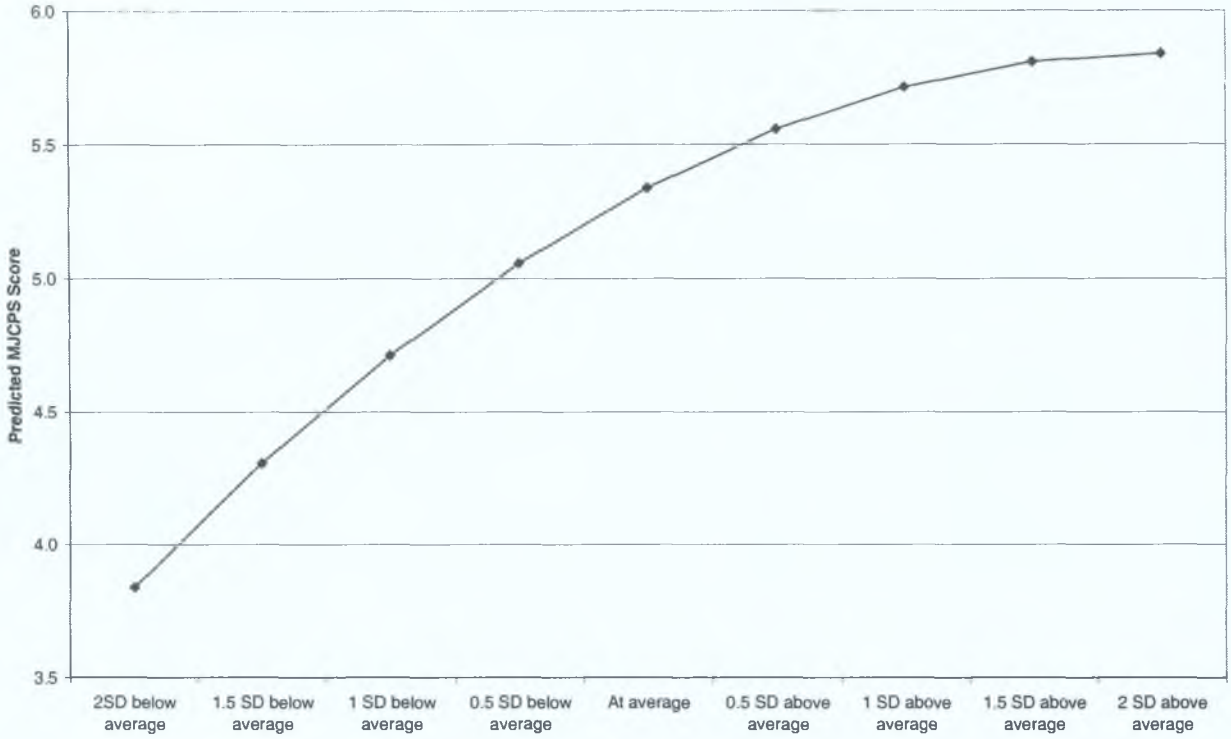
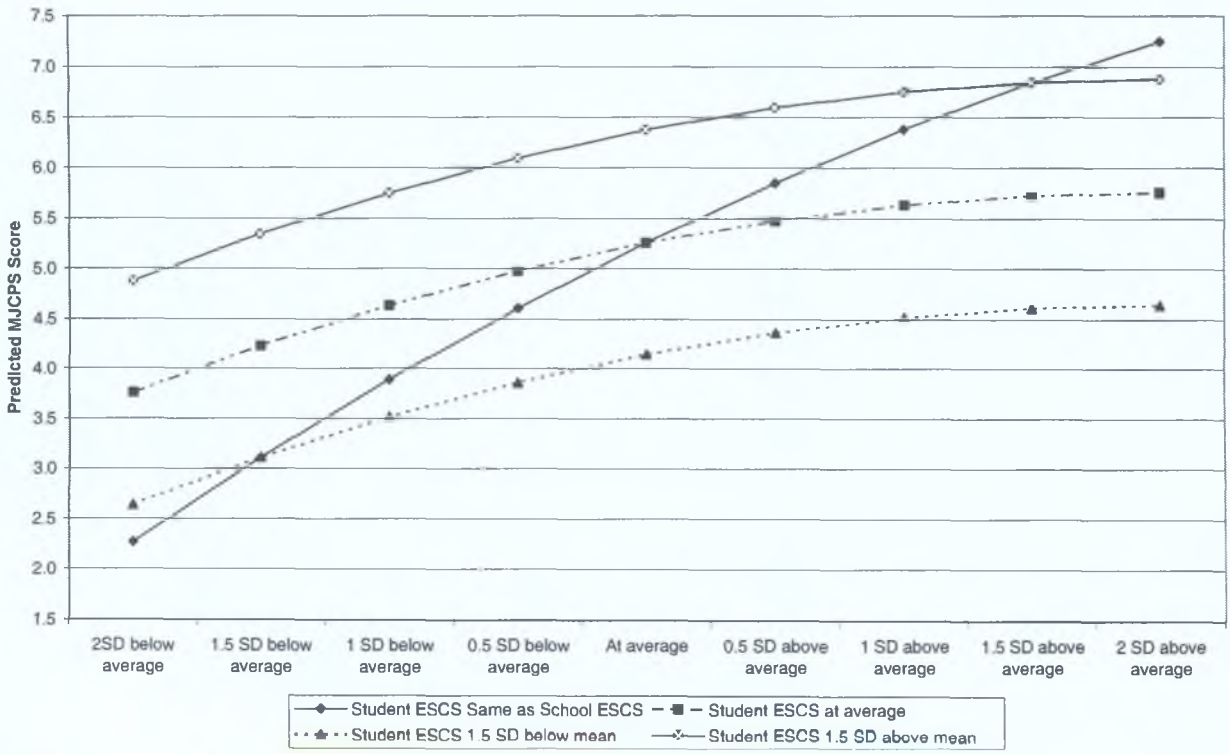


Figure 5.14. Plot of the Relationship Between the Combined Effect of Student and School ESCS and Student Achievement on MJCPS, 2003



5.7.5. Multilevel Models of Achievement on TIMSS 1995 Mathematics

Prior to entering any terms in the model, variance components of the null model were computed to obtain a measure of the total variance that is between schools, which is 43.8%. The mean mathematics score of students included in the model is 533.5, and the standard deviation is 91.3.[42] Tables 5.31 and 5.32 show the parameters for the student-level and school-level variables tested separately. The tables show, for example, that females score about 21 points lower than males, on average, and that there is an estimated 94-point increase associated with a one standard deviation increase in school ESCS. Both of the student-level variables are significant. Of the school-level variables, material resources, school autonomy, teachers' decision-making, and the quality of the school building are not significant. School ESCS explains large portions of the variance (27.4% at the student level and 61.9% at the school level).

Table 5.31. *Achievement on TIMSS 1995 Mathematics: All Student-Level Variables Tested as Separate Models by Addition to the Null Random Intercept Model*

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*	*% student var*	*% sch var*
Gender: female–male	-21.011	4.724	t = -4.448	29	<.001	0.5	0.4
ESCS	8.667	1.747	t =4.963	234	<.001	5.2	10.9

Table 5.32. *Achievement on TIMSS 1995 Mathematics: All School-Level Variables Tested as Separate Models by Addition to the Null Random Intercept Model*

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*	*% student var*	*% sch var*
Single Sex-Mixed Sex	24.152	11.077	t = 2.180	130	.031	1.7	3.8
Sector							
Comm/Comp-Secondary	-38.492	13.968	Ddiff = 8.905	2	.012	3.0	6.8
Vocational-Secondary	-24.359	14.5					
Average ESCS	94.163	7.063	t = 13.332	130	<.001	27.4	61.9
Average Disc. Climate	-50.556	8.930	t = -5.662	130	<.001	8.3	18.7
Material Resources	-2.661	6.035	Ddiff = 0.175	2	.916	0.1	0.3
Missing Mat. Resources	-4.718	14.876					
Student Behaviour	-20.091	4.942	Ddiff = 10.142	2	.006	3.5	7.8
Missing Stud. Behaviour	7.318	12.618					
School Autonomy	10.158	5.789	Ddiff = 4.275	2	.118	1.2	2.6
Missing Sch. Autonomy	-6.368	25.794					
Teachers' Decision-Making	-0.501	5.836	Ddiff = 0.126	2	.939	0.0	0.1
Missing Tch. Decision	-6.364	25.808					
School Building Quality	2.483	5.944	Ddiff = 1.035	2	.596	0.3	0.7
Missing Sch. Building	-16.629	17.836					

[42] These are the means of the five plausible values, unweighted.

Table 5.33 shows the student-level variables entered simultaneously. Both remain highly significant and their parameter estimates are similar to when entered one at a time.

Table 5.33. Achievement on TIMSS 1995 Mathematics: All Student-Level Variables Tested Simultaneously

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	532.206	5.958	t = 89.340	131	<.001
Gender: Female-Male	-20.409	5.621	t = -4.416	28	<.001
ESCS	8.449	1.729	t = 4.887	257	<.001

Table 5.34 shows the parameter estimates for all significant school-level variables entered simultaneously. School ESCS and disciplinary climate retain significance while the other variables are no longer significant. Table 5.35 shows the parameters for school ESCS and disciplinary climate following removal of the non-significant terms.

Table 5.34. Achievement on TIMSS 1995 Mathematics: All School-Level Variables Tested Simultaneously

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	520.198	7.995	t = 65.066	124	<.001
Single Sex-Mixed Sex	0.455	9.096	t = 0.050	124	.961
Sector					
Comm/Comp-Secondary	10.591	11.322	Ddiff = 2.390	2	.303
Vocational-Secondary	-9.333	11.079			
Average ESCS	88.228	7.562	t = 11.667	124	<.001
Average Disc. Climate	-23.981	6.905	t = -3.473	124	.001
Student Behaviour					
Student Behaviour	-2.166	4.345	Ddiff = 0.196	2	.907
Missing Stud. Behaviour	-1.027	8.440			

Table 5.35. Achievement on TIMSS 1995 Mathematics: All Significant School-Level Variables Tested Simultaneously

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	520.953	3.510	t = 148.405	129	<.001
Average ESCS	86.501	7.450	t = 11.611	129	<.001
Average Disc. Climate	-23.921	6.983	t = -3.426	129	<.001

All school- and student-level variables were then entered simultaneously. After checking that they retained significance, curvilinearity of continuous variables was tested. In the case of school disciplinary climate and student ESCS, no evidence of curvilinearity was found; however, school ESCS had a borderline significant curvilinear term which was retained until interactions were tested and then removed before

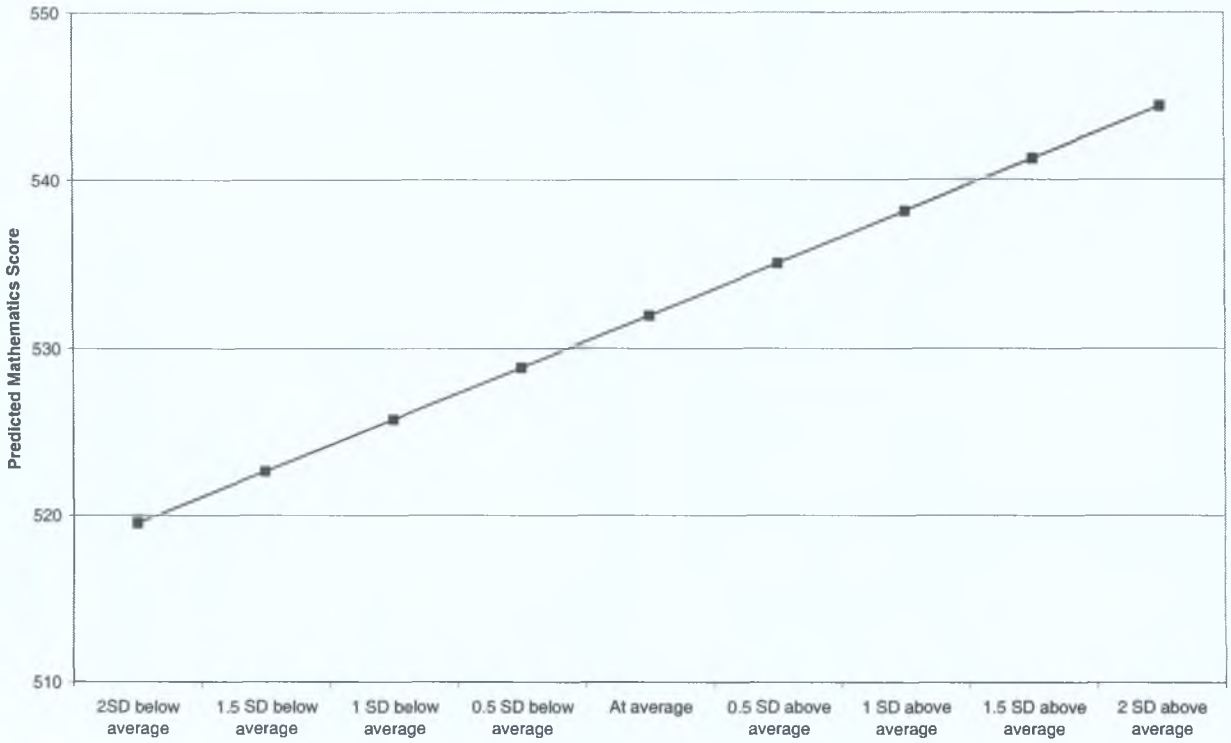
finalising the model. The interaction between gender and student ESCS is not significant, nor is the interaction between gender and school ESCS. As a final check of the model, random components were added to the slopes of the two student-level variables one at a time to see if their effects varied across schools. The effect of student ESCS was found to vary significantly across schools, while the slope associated with student gender was constant. Variation in the ESCS slope is not explained by any of the school-level variables. Factors not accounted for in the model are responsible in the between-school variance in the slope for ESCS. If one takes the square root of the variance of the random slope and adds $\pm$ 1.96 times this to the parameter estimate, one obtains the range of values associated with student gender in 95% of the schools. The square root of the variance is 7.0. The likely range of values associated with a one standard deviation increase in ESCS is therefore -7.5 to 19.9. The final model explains 30.4% of within-school variance, and 66.3% of the variance between schools (or 46.1% of the total variance in achievement). The two school-level variables explain an additional 24.8% of within-school variance, and 55.9% of the variance between schools, which indicates that the majority of achievement variance is explained by school rather than student factors. The final model for PISA 2003 mathematics is shown in Table 5.36.

Table 5.36. Final Model of Achievement on TIMSS 1995 Mathematics

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	531.952	4.383	t = 121.345	129	<.001
Student-Level Variables					
Gender: Female-Male	-20.629	4.161	t = -4.958	35	<.001
ESCS	6.216	1.772	t = 3.507	131	.001
School-Level Variables					
Average ESCS	80.359	7.794	t = 10.311	129	<.001
Average Disc. Climate	-27.710	6.709	t =-4.130	129	<.001
Variance Components					
Intercept variance	1277.146				
ESCS slope variance	48.617				
Level-1 (within-school) variance	4716.141				

The relationship between student ESCS is shown in Figure 5.15. It indicates that the difference between students of high and low ESCS is relatively small, about 25 points (just over a quarter of a standard deviation).

Figure 5.15. Plot of the Relationship Between Student ESCS and Student Achievement on TIMSS 1995 Mathematics



The relationship between 'school' ESCS and achievement is shown in Figure 5.16. The difference in expected achievement on TIMSS mathematics in high- and low-ESCS schools is substantial compared with student ESCS – around 188 points, or two standard deviations. The figure also indicates that the effect of school ESCS on achievement is constant, regardless of the level of ESCS of the school (although there is a slight, borderline significant curvilinear trend, not included in the final model or in Figure 5.16).

Figure 5.17 considers the combined effects of student and 'school' ESCS. This further illustrates the relative disadvantage of students of low ESCS in schools with low average ESCS, and also indicates that, regardless of the ESCS of the student, the relative disadvantage of attending a school of low average ESCS is substantial.

Figure 5.16. Plot of the Relationship Between School Mean ESCS and Student Achievement on TIMSS 1995 Mathematics

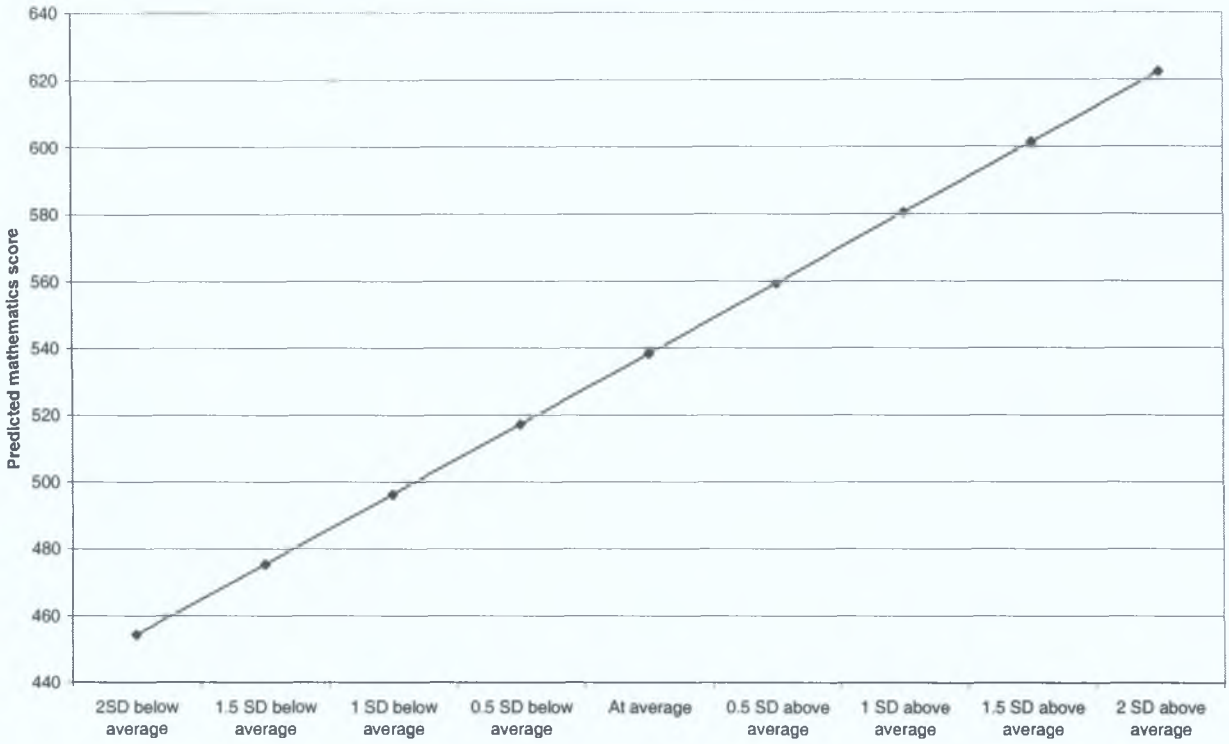
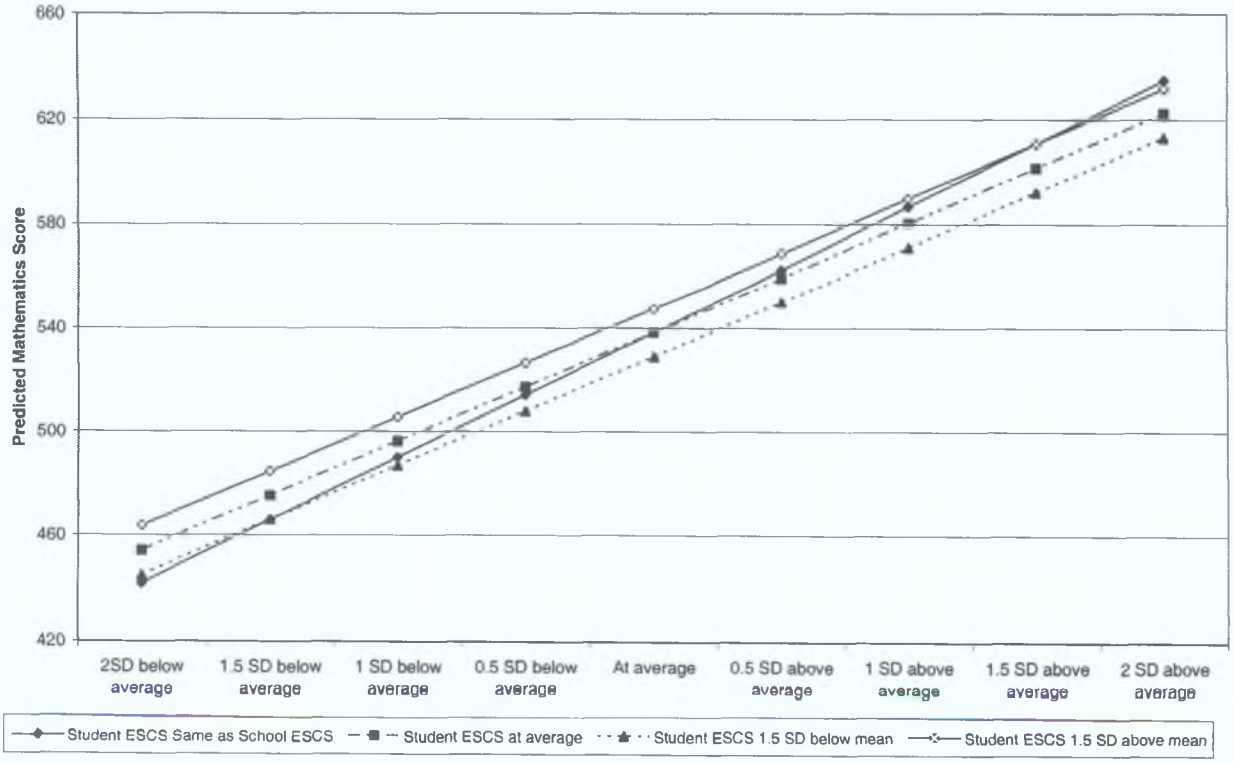


Figure 5.17. Plot of the Relationship Between the Combined Effect of Student and School ESCS and Student Achievement on TIMSS 1995 Mathematics



5.7.6. Multilevel Models of Achievement on Junior Certificate Mathematics for Students Participating in TIMSS 1995

Prior to entering any terms in the model, variance components for the null model were computed to obtain a measure of the total variance that is between schools (61.9%). The mean for the MJCPS of students included in the model is 4.95 and the standard deviation is 2.18. Tables 5.37 and 5.38 show the parameters for the student-level and school-level variables tested separately. As with the model for MJCPS 2000, there is no significant gender difference. The effect for student ESCS is significant. The school-level variables which are significant when tested on their own are ESCS, disciplinary climate, student behaviour, sex composition, and sector, and the effect associated with school ESCS in particular is substantial, with a 1.5 standard deviation increase in MJCPS scores associated with a one standard deviation increase in school ESCS.

Table 5.37. *Achievement on MCJPS, 1996: All Student-Level Variables Tested as Separate Models by Addition to the Null Random Intercept Model*

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*	*% student var*	*% sch var*
Gender: female–male	-0.071	0.079	t = .0.906	2824	.365	0.1	0.2
ESCS	0.242	0.028	t = 8.536	2824	<.001	7.4	10.8

Table 5.38. *Achievement on MJCPS, 1996: All School-Level Variables Tested as Separate Models by Addition to the Null Random Intercept Model*

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*	*% student var*	*% sch var*
Single Sex-Mixed Sex	0.927	0.303	t = 3.062	130	.003	4.2	6.7
Sector							
Comm/Comp-Secondary	-0.920	0.415	Ddiff = 12.03	2	.002	5.6	8.9
Vocational-Secondary	-1.209	0.396					
Average ESCS	2.721	0.197	t = 13.833	130	<.001	39.3	63.1
Average Disc. Climate	-1.607	0.253	t = -6.351	130	<.001	14.3	23.0
Material Resources	0.025	0.166	Ddiff = 0.036	2	.982	0.0	0.0
Missing Mat. Resources	-0.061	0.508					
Student Behaviour	-0.629	0.139	Ddiff = 13.292	2	.001	6.0	9.6
Missing Stud. Behaviour	0.233	0.346					
School Autonomy	0.288	0.167	Ddiff = 3.258	2	.196	1.6	2.5
Missing Sch. Autonomy	-0.042	0.894					
Teachers' Decision-Making	0.079	0.157	Ddiff = 0.249	2	.883	0.1	0.2
Missing Tch. Decision	-0.042	0.894					
School Building Quality	0.111	0.165	Ddiff = 0.827	2	.661	0.4	0.6
Missing Sch. Building	-0.358	0.558					

Table 5.39 shows the student-level variables entered simultaneously. ESCS remains highly significant and gender is not significant in the presence of ESCS; their parameter estimates are similar to when entered one at a time.

Table 5.39. Achievement on MJCPS, 1996: All Student-Level Variables Tested Simultaneously

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	4.633	0.155	t = 29.827	131	<.001
Gender: Female-Male	-0.055	0.075	t = -0.738	2823	.461
ESCS	0.242	0.028	t = 8.522	2823	<.001

Table 5.40 shows the parameter estimates for all significant school-level variables entered simultaneously. School ESCS and disciplinary climate retain their significance. School sector is borderline significant, and the direction of the parameter estimates indicate that, after adjusting for the other variables in the model, students in vocational schools outperform students in secondary and community/comprehensive schools. The other two variables (sex composition and student behaviour) are not significant. Table 5.41 shows the parameters for school ESCS, disciplinary climate and school sector following removal of the non-significant terms. For now, school sector, which is borderline significant, is retained, since these variables have not yet been tested together with the student-level variables.

Table 5.40. Achievement on MJCPS, 1996: All School-Level Variables Tested Simultaneously

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	4.456	0.193	t = 23.034	124	<.001
Single Sex-Mixed Sex	0.236	0.220	t = 1.076	124	.285
Sector					
Comm/Comp-Secondary	-0.316	0.263	Ddiff = 5.243	2	.073
Vocational-Secondary	0.371	0.292			
Average ESCS	2.444	0.214	t = 11.396	124	<.001
Average Disc. Climate	-0.848	0.170	t = -4.997	124	<.001
Student Behaviour					
Student Behaviour	-0.110	0.071	Ddiff = 1.076	2	.584
Missing Stud. Behaviour	-0.050	0.217			

Table 5.41. Achievement on MJCPS, 1996: All Significant School-Level Variables Tested Simultaneously

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	4.623	0.109	t = 42.366	127	<.001
Sector					
Comm/Comp-Secondary	-0.496	0.229	Ddiff = 5.521	2	.063
Vocational-Secondary	0.179	0.250			
Average ESCS	2.496	0.212	t = 11.758	127	<.001
Average Disc. Climate	-0.855	0.175	t = -4.877	127	<.001

The curvilinearity of continuous explanatory variables was then tested. There is no evidence of curvilinearity for student ESCS or school disciplinary climate. However, school ESCS shows a significant curvilinear trend. The interaction between gender and student SES is not significant. A test of the constancy of slope variation across schools for the student variables was made by introducing an error term to the slope for gender and ESCS one at a time in the model, and the results indicate that the slopes are constant across schools for both ESCS and gender. There is no cross-level interaction between gender and school ESCS. Gender was then removed from the model, since it is not itself significant; nor does it contribute to any interactions. Table 5.42 shows the final model for achievement on MJCPS which explains 45.8% of within-school variance, and 72.5% of the variance between schools (62.3% of the total variance in achievement). The school-level variables explain an additional 38.4% of within-school variance, and 61.7% of the variance between schools.

Table 5.42. Final Model of Achievement on MJCPS 1996

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	4.653	0.114	t = 40.748	126	<.001
Student-Level Variables					
ESCS	0.214	0.029	t = 7.379	2818	<.001
School-Level Variables					
Sector					
Comm/Comp-Secondary	0.271	0.239	Ddiff = 6.482	2	.039
Vocational-Secondary	-0.466	0.228			
ESCS Parameters					
Average ESCS	2.300	0.182			
Average ESCS Squared	-0.704	0.205	t = -3.433	126	.001
Average Disc. Climate	-0.889	0.162	t = -5.491	126	<.001
Variance Components					
Intercept variance	0.852				
Level-1 (within-school) variance	1.890				

The relationship between student ESCS and achievement on the MJCPS is shown in Figure 5.18. There is a relatively small difference in the expected MJCPS scores, of just a quarter of a standard deviation, between students with an ESCS two standard deviations above and below the mean. The curvilinear nature of the relationship between 'school' ESCS and achievement is shown in Figure 5.19. The gradient has a much gentler curve than those associated with the four PISA models.

Figure 5.18. Plot of the Relationship Between Student ESCS and Student Achievement on MJCPS, 1996

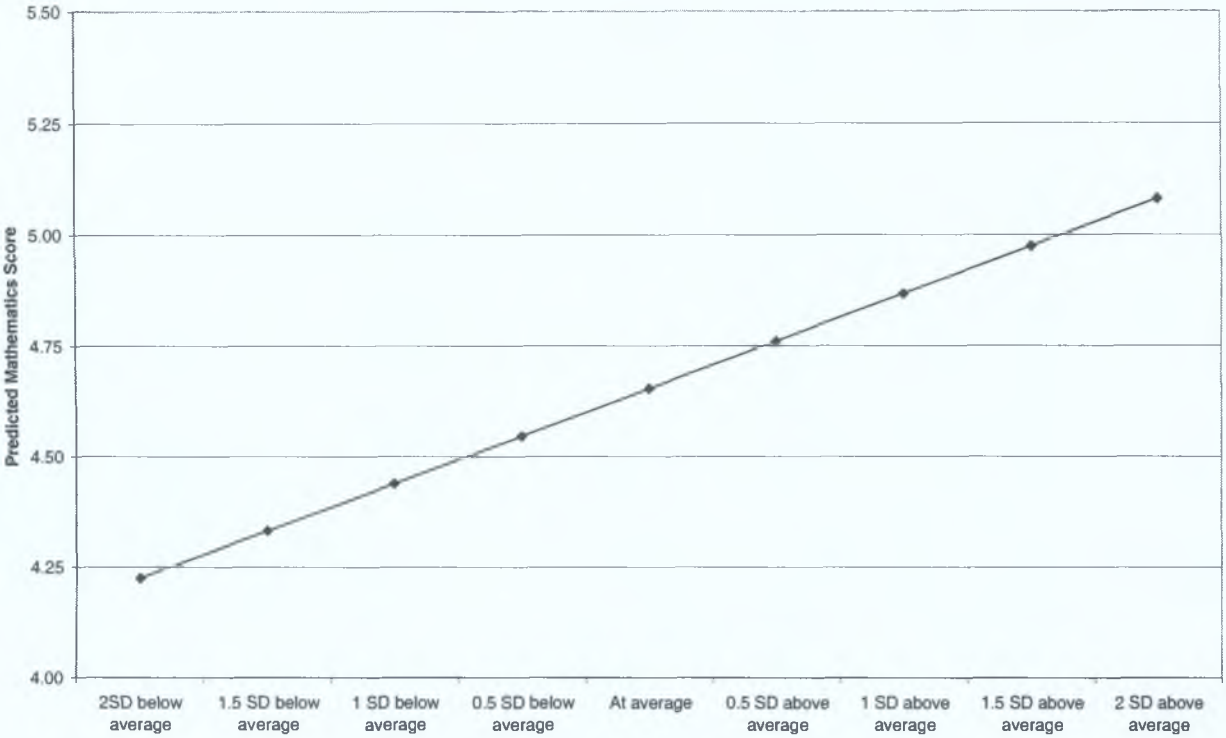
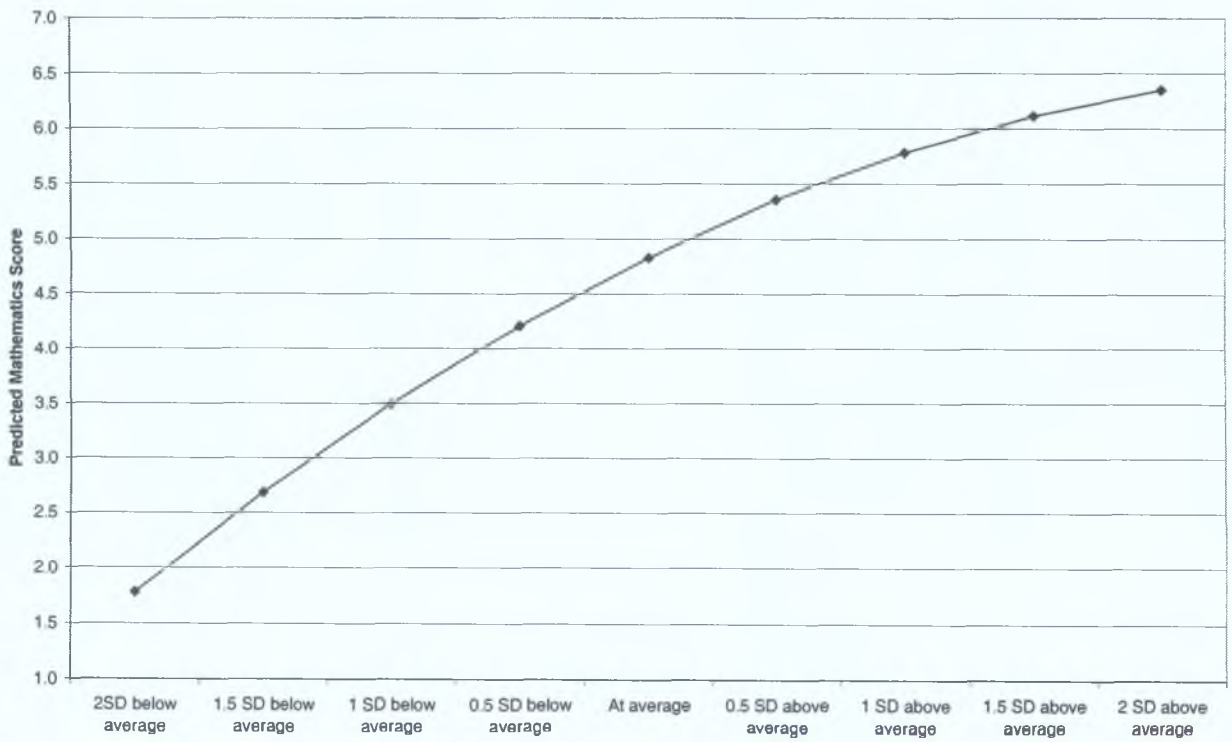
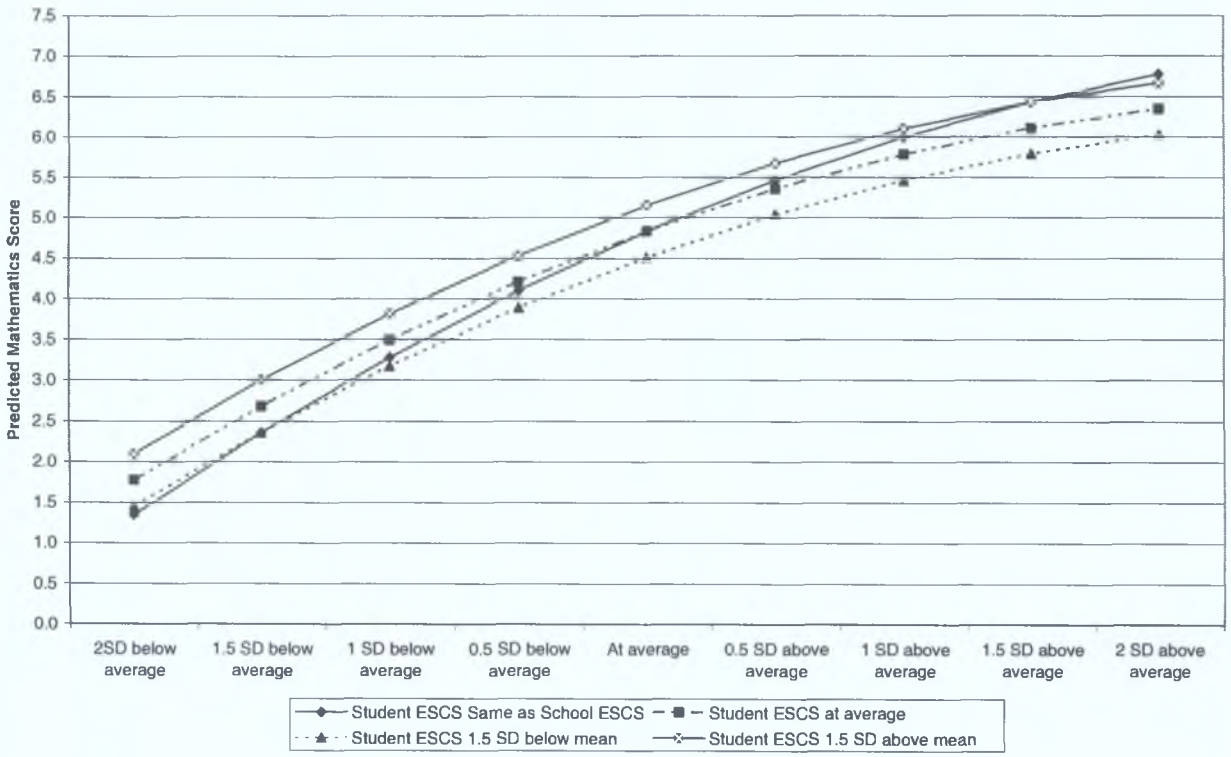


Figure 5.19. Plot of the Relationship Between School Mean ESCS and Student Achievement on MJCPS, 1996



The combined 'school' and student effects are shown in Figure 5.20, which indicates that the predicted difference in MJCPS scores of students with an ESCS that is two standard deviations below the average, and who are in schools with a mean ESCS that is two standard deviations below average, compared with students with an ESCS that is two standard deviations above the average, and who are in schools with a mean ESCS that is two standard deviations above the average is substantial; around two-and-a-half standard deviations.

Figure 5.20. Plot of the Relationship Between the Combined Effect of Student and School ESCS and Student Achievement on MJCPS, 1996



5.7.7. Exploration of the Curvilinearity of the Social Context Effect in 2003

A curvilinear trend in the school-level ESCS measure was found in all of the multilevel models presented with the exception of TIMSS 1995 mathematics. Although not a focus of the research questions addressed in this chapter, it stands in contrast to the models reviewed in Chapter 2, where no evidence of curvilinearity was reported. Since the measure of student social background used in the models in this chapter was a composite of a number of aspects of social background, I wanted to explore further whether the observed curvilinear nature had to do with the manner in which school social intake was measured. Therefore, I used, as an alternative to school ESCS, a

weighted average of the percentage of students entitled to a fee waiver for the Junior Certificate, and re-computed the parameter estimates for PISA 2003 mathematics and MJCPS 2003 (cf. Tables 5.24 and 5.30). The fee waiver variable is similar to that used by Sofroniou et al. (in preparation) (described in Chapter 2) and this variable was also used in the models of achievement in PISA 2003 (Cosgrove et al., 2005). Table 5.43 shows the model for PISA 2003 mathematics with fee waiver rather than ESCS. The squared term for fee waiver does not nearly approach significance in the model (p = .448) so only the original term is included. The slope for gender is significant in this model also. The model explains 18.8% of variance within schools, and 69.4% between schools. Slightly less of the between-school variance is explained by the model in Table 5.43 than the PISA 2003 model that uses school ESCS (which explains 20.0% of variance within schools and 77.2% between schools).

Table 5.44 shows the final model of MJCPS 2003, again with fee waiver instead of school ESCS. The squared term is once again not significant (p = .910). The variance explained by the final model is 25.6% within schools and 84.9% between schools; similar to the model which uses school ESCS (i.e., 26.0% within schools and 86.7% between schools). The effects associated with school sector and school building quality are stronger in the model which uses fee waiver compared with the model which uses school average ESCS. Contrary to Sofroniou et al.'s findings, there is no cross-level interaction with gender in either model.

Table 5.43. Model of Achievement on PISA 2003 Mathematics With Examination of Curvilinearity of Junior Certificate Fee Waiver

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	500.780	2.918	t = 171.605	140	<.001
Student-Level Variables					
Gender: Female-Male	-15.967	1.936	t = -4.121	141	<.001
ESCS	29.506	1.936	t = 15.238	394	<.001
School-Level Variables					
Fee Waiver	-0.937	0.151	t = -6.194	140	<.001
Variance Components					
Intercept variance	412.725				
Gender slope variance	455.142				
Level-1 (within-school) variance	5177.369				

Table 5.44. Model of Achievement on MJCPS 2003 With Examination of Curvilinearity of Junior Certificate Fee Waiver

	Parameter	*SE*	*Test Statistic*	*df*	*p–value*
Intercept	5.272	0.075	t = 70.149	135	<.001
Student-Level Variables					
ESCS intercept	0.883	0.054	t = 16.348	2284	<.001
School-Level Variables					
Sector					
Comm/Comp-Secondary	0.397	0.136	Ddiff = 9.666	2	.008
Vocational-Secondary	-0.084	0.151			
Fee Waiver	-0.035	0.004	t = -8.315	135	<.001
Average Disc. Climate	0.830	0.131	t = 6.326	135	<.001
School Building Quality					
School Building Quality	-0.174	0.054	Ddiff = 14.025	2	<.001
Missing Sch. Building	-0.323	0.29			
Variance Components					
Intercept variance	0.134				
Level-1 (within-school) variance	4.143				

5.7.8. A Comparison of the Multilevel Models

In Sections 5.7.1 to 5.7.6, six multilevel models of student achievement were presented. The main characteristics of these are summarised here. All models presented in this chapter (i.e., in Sections 5.7.1 to 5.7.7) are then used in discussion of the research questions described in Section 5.5.

In the six models considered initially, a common set of variables was examined: student gender and student ESCS, and at the school level, sector, sex composition, school average ESCS, school disciplinary climate, student behaviour, school material resources, quality of the school building, school autonomy, and teacher participation in decision-making. Sample stratum was included in the four PISA models to account for any variance in achievement arising from the sample design; this variable was not required in the TIMSS models.

In all of the models, school and student ESCS are significant, giving strong support for the presence of an effect for social intake. This 'social context effect' is particularly strong in the TIMSS models, along with a weaker association between student ESCS and achievement. In five of the six models, a significant curvilinear trend was found for social intake whereby the effect is weaker at higher levels of school ESCS. The TIMSS model suggested a linear social context effect, while the slope for MJCPS 1996 is

gentler than the slope associated with the four PISA models, where the achievement gradient flattened out at values around one standard deviation above average ESCS. Further, none of the models required an interaction term for student gender and student ESCS, or for student gender and school ESCS, which suggests that a student's social background operates in a similar manner regardless of gender, as does the school social context effect.

The final model for PISA 2000 reading contains just four variables: gender (where females outperform males), student ESCS, school ESCS, and school disciplinary climate. The model for EJCPS 2000 had a borderline significant term for school sex composition ($p = .071$), whereby students in single sex schools outperform those somewhat in mixed sex schools; this was dropped from the final model. The model for EJCPS 2000 also required a random term for the slope associated with student gender. Some of the slope variance associated with student gender is explained by the quality of the school building, although the slope still varies significantly across schools.

The final model for PISA 2003 mathematics required just three variables – student ESCS, school ESCS and gender (whereby males outperform females). The slope for gender also varies significantly in this model, whereby the gender difference, favouring boys, is expected to range from 55 points (in favour of boys) to 11 points (in favour of girls). The variance in the slope for gender is not explained by any of the school-level variables. The final model for MJCPS 2003 contains student ESCS, school ESCS, quality of the school building, and school sector, and school disciplinary climate. The gender difference is not significant. The parameters for school sector suggest that, after adjusting for the other variables, students in community/comprehensive schools outperform their counterparts in secondary and vocational schools.

The final model for TIMSS 1995 includes gender (where boys outperform girls), student ESCS, school ESCS and school disciplinary climate. The slope for student ESCS varies across schools (the effect of which, as noted, is relatively small), and this variation is not explained by any of the school-level variables. The final model for MJCPS 1996 does not include gender (consistent with the model for MJCPS 2003), and contains student ESCS, school ESCS, school disciplinary climate, and school sector. The parameters for school sector suggest that, after adjusting for the other variables,

students in vocational schools achieved lower average scores than those in secondary schools; the parameter for community/comprehensive is positive, indicating slightly higher performance than in secondary schools.

Table 5.45 compares the variance components of the six models presented in Sections 5.7.1 to 5.7.6. The models explain between 29% and 62% of the total variance in achievement; between 66% and 87% of the variance between schools; and between 20% and 46% of the variance within schools.

The percentage of total variance that is between schools is similar for the models of PISA 2000 reading and EJCPS (18% compared to 21%). A comparison of PISA 2003 mathematics and MJCPS 2003 indicates that the between-school variance is a little higher for MJCPS 2003 (15% compared to 20%); and the same pattern is evident when one compares TIMSS and MJCPS 2003 (44% compared with 62%).

The models of EJCPS 2000 and MJCPS 2003 explained more of the within-school variance than the models of PISA 2000 and PISA 2003. The effect associated with student gender varies across schools in the case of EJCPS, and in the case of PISA 2000 reading, school ESCS explains slightly more of the variance between schools; in other respects, the models of English/reading are quite similar. The models for PISA 2003 and MJCPS 2003 are more different to one another. The model for PISA 2003 mathematics does not contain any school-level variables apart from school ESCS. In contrast, the model for MJCPS 2003 includes school ESCS as well as school sector, disciplinary climate, and school building quality. Hence, the model for MJCPS 2003 explains comparatively more achievement variance, both within and between schools. Student and school ESCS explain slightly more of the achievement variance in the model for MJCPS 2003 (particularly within schools).

The model for MJCPS 1996 explained the most achievement variance out of all six models (62%), which is also higher than the total explained variance for TIMSS 1995 (46%). 'School' ESCS explains more of the within-school variance in MJCPS 1996; this seems to account for most of the difference in the variance explained of the two models. However, while the model for TIMSS 1995 included gender, this was not the

case with MJCPS 1996 (consistent with MJCPS 2000); further, school sector explains a significant amount of achievement variance in MJCPS 1996, but not in TIMSS 1995.

Table 5.45. Comparison of Variance Components of the Six Models

		Variance Explained By:						
	Null Model Variance	*Student ESCS*	*School ESCS*	*Student & School ESCS*	*All Student*	*All School*	*Unique variance associated with school/class practice*	*Final Model*
PISA 2000 Reading								
Within schools	81.6	12.7	12.8	18.2	15.4	13.7	2.3	21.4
Between schools	18.4	36.9	65.2	64.5	47.9	70.3	3.4	76.7
Total	100	17.2	22.5	26.7	21.4	24.1	2.5	31.6
EJCPS: 2000 Cohort								
Within schools	79.4	13.7	13.2	19.1	19.6	14.1	1.0	25.4
Between schools	20.6	36.5	60.8	60.9	54.0	64.9	4.3	79.6
Total	100	18.4	23.0	27.7	26.7	24.6	1.7	36.6
PISA 2003 Mathematics								
Within schools	85.2	15.4	12.5	19.2	16.2	12.5	0.0	20.0
Between schools	14.8	53.9	78.3	76.6	55.0	78.3	0.0	77.2
Total	100	21.1	22.2	27.7	21.9	22.2	0.0	28.5
MJCPS: 2003 Cohort								
Within schools	79.8	18.5	16.1	24.2	18.5	18.6	1.8	26.0
Between schools	20.2	52.7	75.3	78.0	52.7	87.1	8.7	86.7
Total	100	25.4	28.1	35.1	25.4	32.4	3.2	38.3
TIMSS 1995 Mathematics								
Within schools	56.2	5.2	27.4	27.8	5.6	29.0	0.9	30.4
Between schools	43.8	10.9	61.9	61.8	10.4	65.5	3.5	66.3
Total	100	7.7	42.5	42.7	7.7	45.0	2.0	46.1
MJCPS: 1996 Cohort								
Within schools	38.1	7.4	40.3	41.0	7.4	45.1	4.8	45.8
Between schools	61.9	10.8	64.8	64.6	10.8	72.5	7.9	72.5
Total	100	9.5	55.5	55.6	9.5	62.1	6.7	62.3

5.7.9. How the Analyses Address the Research Questions

In Section 5.6, several research questions were posed. These are revisited in this section with a description of how the analyses have addressed them, and whether hypotheses have been supported.

5.7.9.1. Test Content

First, since the domain mathematics is more school-dependent than English/reading, mathematics achievement should be more sensitive to school-level effects. This hypothesis received some support when the models of English/reading are compared with those of mathematics, where, of nine variables relating to school resources and

climate, just one (disciplinary climate) was in the final model of EJCPS 2000 and PISA 2000 reading. In contrast, the two models of Junior Certificate mathematics required additional school-level variables. That said, the model of PISA mathematics did not require any variables at the school level other than school ESCS.

I also argued, because the test-curriculum rating project suggested notable disparities between PISA mathematics and Junior Certificate mathematics, both in the concepts assessed, and in the manner in which problems are contextualised, and because TIMSS mathematics is intended to be only somewhat compatible with national mathematics curricula, that Junior Certificate mathematics would be more sensitive than both PISA mathematics and TIMSS mathematics to school-level effects. This hypothesis received support, albeit that the range of school-level variables was restricted, and of somewhat limited quality (perhaps inevitable in a cross-sectional survey design whose composites are largely based on the opinions of students and principals). The final model for Junior Certificate mathematics in 2003 had three significant school-level variables other than ESCS and these explained 1.8% and 8.7% of additional variance at the student and school levels, respectively, over and above school ESCS. The school-level variables (other than ESCS) in the model for MJCPS 1996 also explained proportionately more of the within-school (4.8%) and between-school (7.9%) variance compared with TIMSS 1995 (0.9% and 3.5%, respectively).

I further hypothesised, regarding curriculum sensitivity, that due to similarities in the reading processes assessed in PISA reading and Junior Certificate English (and the less school-dependent nature of reading), the explanatory models for both of these would be very similar. Broadly speaking, this received support; however the slope associated with gender for EJCPS is associated with school building quality and suggests that this and other school-level variables which have not been considered in the models may be mediating achievement on EJCPS but not on PISA 2000 reading.

5.7.9.2. Impact of Social Intake

First, I hypothesised that the association between school-level SES and achievement would be strong in all models examined. Although strong support for a social context effect was found in all models examined, particularly TIMSS 1995 and MJCPS 1996, the effect was curvilinear rather than linear in five of the six models which used school

ESCS, with a tapering off of the strength of the effect at higher values of school ESCS. In contrast, the two models of achievement in 2003 which used Junior Certificate fee waiver as the school social context measure had a linear association with achievement. This suggests that income-related measures may be more likely to have a linear association with achievement than composite measures which incorporate home educational climate and material possessions. Thus while the models that use school-level ESCS support Willms' (2002) hypothesis of diminishing returns, the models that use fee waiver do not, and the latter finding is consistent with Sofroniou et al. (in preparation).

Second, I argued that the effects associated with social intake would be weaker in the models of mathematics compared with English/reading. Comparing the models associated with 2000 and 2003 which use school ESCS as the social context measures, this hypothesis did not receive support; in fact, regardless of whether one considers achievement on PISA or the Junior Certificate, school-level ESCS explains proportionately more of the between-school variance for mathematics, over and above that of student ESCS. This contrasts with Sofroniou et al.'s (in preparation) models and suggests that the manner in which social context is measured (i.e., whether income-related or a composite) may be relevant in considering this issue.

Third, I argued that, if students are clustered within classrooms on the basis of social background, as well as on the basis of ability, the strength of the social context effect for TIMSS 1995 and MJCPS 1996 would be stronger than the social context effect associated with models for 2000 and 2003. This hypothesis received strong support. It suggests that estimates of the social context effect should take account of the sample design. One could hypothesise that the large differences in the social context effects between the TIMSS and PISA models might relate to the extent to which students of similar SES are clustered within schools/classes. But a comparison of proportion of total variance in ESCS that is between schools in TIMSS 1995, PISA 2000 and PISA 2003 suggests that this is not the case. The intra-cluster correlations, respectively, are 22.3%, 18.4%, and 20.7%. The differences could be due to clustering of students within classes based on factors other than ESCS, but nonetheless which mediate the relationship with ESCS and achievement, or again, the fact that different, although comparable, components, make up the ESCS measure in TIMSS and PISA.

Finally, although not a focus of the present research, it is worth noting that although a cross-level interaction between the social context and gender was expected (whereby the slope would be steeper for males) none was found, either for ESCS or for fee waiver. Perhaps the student-level SES measure is also of relevance here. Unfortunately, PISA and TIMSS do not have a reliable income-related measure such as medical card status so this possibility cannot be tested using the datasets considered here.

5.8. Conclusion

This chapter attempted to address two questions: First, what does PISA tell us about the equity of achievement outcomes in Ireland? Second, what does PISA tell us about the determinants of achievement in Ireland? These arose from concerns identified in Chapters 1 and 2 that the sample design and the subject area pertaining to the achievement measure and its alignment to the curriculum may impact on both of these questions. These concerns are considered in light of the widespread discussions of 'educational equity' of the OECD and the appearance of these themes in Irish media and government commentary on PISA, as well as the possibility that conclusions one might draw about the relative impact of school/class variables in explanatory analyses may vary depending on the sample design and the achievement measure. It was also noted that no recent research has addressed these themes directly. Unfortunately, comparisons across the various datasets used in the analyses reported in this chapter are hampered by differences in their design; therefore the results presented here should be taken as *prima facie* evidence of their importance rather than a definitive quantification of their impact.

A comparison of the variance components associated with TIMSS 1995 mathematics and PISA 2000 suggests that in some countries including Ireland, the manner in which achievement variance is partitioned between and within schools varies considerably, where intact-class sampling results in much higher between-school achievement variance than random within-school sampling. Martin et al. (2000b) do not justify selecting grade 8 students only as basis for the indicator of between-school variance in the TIMSS publication on school effects on student achievement (it would have been possible to model the achievements of students from two intact classes per school within a two-level or a three-level multilevel model), and the re-analysis of variance components associated with TIMSS mathematics for Ireland demonstrates that such a

choice makes between-'school' variance appear comparatively high in Ireland. This exercise also confirms research that suggests that much of the achievement variance between schools may be attributed to differences between classrooms and/or streaming practices (Kellaghan et al., 1979; Madaus et al., 1976, 1979; OECD, 2004a; Smyth, 1999;).

A comparison of the variance components associated with TIMSS 1995 and the Junior Certificate mathematics scores of students who participated in TIMSS and who sat the Junior Certificate in 1996 or 1997 indicates that between-cluster variance is higher on the Junior Certificate measure, which provides support for the argument that achievement on a curriculum-sensitive and school-dependent test is more sensitive to school and class effects than a somewhat more generic measure such as TIMSS; however, this increase may be partly attributable to the fact that the TIMSS measure was taken earlier than the Junior Certificate. It's possible that an additional year of schooling may have caused some of the inflation in between-school differences.

A comparison of the variance components associated with PISA 2000 reading and EJCPS indicate that there is practically no difference in the variance components for the two measures, whether all students, or third year/grade 9 students only are considered. This finding lends support to the argument advanced in Chapter 2 that PISA reading and Junior Certificate English share considerable commonalities in terms of the reading processes tested. The finding could also have arisen due to the possibility that skills associated with reading are more generic and that English is a less school-dependent subject than others, such as mathematics or science, particularly by the age of 15.

It was found that a higher percentage of between-school variance is associated with Junior Certificate mathematics than with PISA 2003 mathematics, regardless of whether all students, or just students in third year/grade 9, are considered. This provides some support for the argument that, since PISA mathematics represents a considerable departure from Junior Certificate mathematics, it is less sensitive to school effects.

Several multilevel models were presented. Using data from PISA 2000, PISA 2003 and TIMSS 1995, they were designed to address a specific set of research questions relating

to the nature of the measure, sample design, and the impact of social intake and school/class variables on achievement.

The findings on curriculum sensitivity suggest that it is indeed important to consider *both* the extent to which a test is designed to assess the curriculum *and* the degree to which the subject area tested is school independent or not, since these factors can result in different conclusions being drawn about the nature and extent of school-level effects. For example, in the model for PISA 2003 mathematics, one can conclude that the school has little if any bearing on the achievements of students; student gender and school and student ESCS are the only variables in the model. In contrast, in the model for MJCPS 2003, school-level variables explained an additional 3.2% of the total variance (1.8% at the student level and 8.7% at the school level), over and above student and school ESCS. This is also true of the TIMSS 1995 and MJCPS 1996 models, where school-level variables explained an additional 6.7% of total variance over and above student and school ESCS in MJCPS 1996 compared to just 2.0% for TIMSS mathematics. The additional explained variance is in fact likely to be an underestimate in all cases since school and student ESCS have not been adjusted for student intake.

It would appear that English/reading is a relatively school-independent skill, particularly so, perhaps, at the age at which PISA is administered. At any rate, school-level variables in the models of PISA 2000 reading and EJJCPS explained only about 2% of the total variance over and above student and school ESCS.

The lack of a significant interaction between gender and school social background in any of the models contradicts the results of multilevel models presented by Sofroniou et al. (in preparation), where the detriments associated with low SES were generally greater for boys than girls. In making comparisons between these models, it was noted that the social background measure in the modeling of TIMSS and PISA data were composites, comprising indicators of parental education, occupation, home educational resources, and material possessions; the models in Sofroniou et al. used entitlement to medical card/examination fee waiver as the only measure of SES, and in this sense, it is a measure of economic deprivation rather than a measure of the wider social background used here.

However, the re-analysis of the PISA 2003 and MJCPS 2003 models which used fee waiver rather than school ESCS *also* failed to detect a significant cross-level interaction. One could further hypothesise that not only is the nature of the social intake measure relevant to detecting cross-level interactions associated with gender, the student-level measure is also relevant. It is quite possible that, had an income-based student measure of SES such as student medical card possession been used in the models presented here, rather than composite ESCS, that a cross-level interaction consistent with Sofroniou et al. would have been detected. Unfortunately, these data are not available for PISA or TIMSS students.

Another striking feature of the models presented in this chapter is that the school social context effect arising from ESCS is curvilinear for all measures (significant in all cases except for TIMSS mathematics, where it is borderline significant). In all cases, the curvilinearity operates in a similar manner: the social context effect is weaker at relatively high levels of school ESCS than at relatively low levels of ESCS; the plotted values tend to flatten out as average school ESCS approaches around one standard deviation above the average. There is some evidence, when one compares the PISA 2003 and MJCPS 2003 models which use school ESCS and those which use Junior Certificate fee waiver, that whether or not the school context effect is curvilinear is associated with the manner in which social context is measured (i.e., an aggregate of income versus a composite aggregate based on parental occupation, education, and home material and educational possessions).

A comparison of the explanatory power of student and school ESCS in the TIMSS and PISA models indicates that, in TIMSS, student ESCS does not predict substantial proportions of achievement variance, either on its own or in conjunction with other variables, whereas school ESCS exerts a strong effect on both student achievement and school average achievement, both on its own and with other variables. In contrast, both student and school ESCS exert substantial effects on student achievement, both on their own and in conjunction with other variables in the PISA models. This finding is unrelated to the extent to which students of similar SES are clustered within schools/classes, since the between-school variance associated with student ESCS is about the same across all three datasets. It is possible that differences could be due to clustering of students within classes based on factors other than, but associated with,

ESCS, or to differences in the ESCS measures used in TIMSS and PISA. This finding merits more in-depth investigation, since the impact of social intake on achievement is of current and enduring policy relevance. Notwithstanding the limitations in comparisons that may be drawn due to differences in the nature of the ESCS measures used in PISA and TIMSS and other differences in the survey design, test content, etc., the TIMSS models would suggest that targeted interventions aimed at alleviating low achievement associated with socioeconomic disadvantage should be made with reference to school ESCS since students in low-ESCS schools, regardless of individual ESCS, have similarly low predicted scores, while the PISA models suggest that targeted interventions need to take both student and school ESCS into account, unless it is the case that low-ESCS students are largely clustered within low-ESCS schools (in which case the target group can be defined by school ESCS).

A number of limitations with the analyses presented in this chapter should be noted. First, no intake measure has been included. The likely result of this is that the effect of ESCS at student and school levels has been inflated and much of the achievement variance within schools remains unexplained. Had they been available, further improvements could have been made to the adjustment of student outcomes by the inclusion of, for example, information on the primary school attended by students; how students selected (or were selected into) the post-primary school they were currently in; and whether students had switched schools after starting post-primary (see Goldstein, 2004).

Second, the models only provide a broad description of the relationships between SES at the student and school levels. They reveal the existence of relationships between school and student variables but they cannot be used to make causal inferences or predictions. A third limitation of the models is that the measures of school inputs or processes have relatively little explanatory power (see Smithers, 2004). This may be because the measures are largely based on principals' and students' opinions rather than on quantifiable attribute and/or because a cross-sectional design may not be the optimal manner in which to assess school effects (Goldstein, 2004). The OECD (2005c) has provided a tentative ranking of the robustness of 'causal' variables in its recent analysis of school factors and has accorded higher robustness to school structural features, and

the SES of students and schools, and lower robustness to school and teaching processes and school climate. School resources are placed in the middle of this rating.

Fourth, the relationships between measures may be different at the extremes; this possibility has not been tested in the models. For example, O'Donoghue, Thomas, Goldstein & Knight (1997) examined progress from GCSE to A Level and showed that the relative progress of both high and low achievers was significantly different to the relative progress of the average student. The fitting of a spline function to the scores of higher achievers improved the fit of their model. Further, the obtained samples for PISA 2000 and 2003 were shown in Chapter 2 to be significantly biased, with lower student response rates for some subgroups. Therefore, just as the overall achievement results (and their distributions) may be unreliable, the results of the multilevel models may be unreliable for estimating the effects of very low achievers, since many of these students did not even participate. [Given the argument of Monseur & Wu (2002) that student non-response adjustments are more efficient in samples with higher between-school variance, the TIMSS achieved sample is unlikely to suffer from the same degree of bias as the PISA samples, particularly given that the weighted student response rate for grade 8 (91%) in TIMSS (Beaton et al., 1996a) was slightly higher than for both PISA 2000 and PISA 2003.]

Finally, Goldstein (2000) has drawn our attention to the fact that schools interact with one another, but the models assume that schools are non-interacting entities. To this one could add that not only are students clustered within schools more similar to one another, so are schools within communities. Willms (2002) notes that there has been relatively little research on the contextual effects of communities other than schools or classrooms. Therefore, a further extension of this work could examine community effects in addition to those associated with schools, classes and students within a three-level model. It may be the case that segregation on the basis of achievement, social background, or other attributes, is much more marked at the level of communities than the level of schools. Of course, what constitutes a 'community' is a much more complex matter than what constitutes a school.

Although not a focus of the analyses presented in this chapter, the inconsistent pattern of gender differences in the models of MJCPS compared with PISA 2003 mathematics

and TIMSS mathematics is noteworthy. There is no significant gender difference on MJCPS, while gender difference favouring males were found in both PISA 2003 and TIMSS 1995. Sofroniou et al. (2002) also found that there was no gender difference in performance on Junior Certificate mathematics for students participating in PISA 2000. There are several possible reasons for this. It was noted (Cosgrove et al., 2005) that, of the four PISA 2003 mathematics subscales, the largest gender differences were observed for the Space & Shape subscale, where Irish males outperformed females by about 26 points – just over a quarter of a standard deviation. There is consistent evidence that males, on average, tend to be better at spatial reasoning tasks than females (Collaer & Nelson, 2002; Voyer, Voyer & Bryden, 1995; Watson & Kimura, 1991); this may in part explain the performance gap on the Space & Shape subscale, and the lack of a gender difference on MJCPS, where items assessing spatial reasoning is not tested in the same way as in PISA. However, the TIMSS 1995 mathematics test did not have a large spatial reasoning component either (Adams & Gonzalez, 1996). Here, it is possible that item format may have contributed to this difference. There is some evidence for gender differences in performance relating to item format. In a study of Irish 15-year-old students, comparisons of performance on multiple-choice (standardised test) and free-response (public examinations) measures of mathematics and English indicated that males performed better than females on multiple-choice tests, and females performed better than males on free-response tests, with the effect associated with item format apparent across both subject domains. (See also Zhang, Wilson, & Manon, 1999; Wilson & Zhang, 1998 for reviews on the topic.) The TIMSS 1995 mathematics test comprised 75.6% of multiple-choice items.

In conclusion, in the absence of strong measures of school/class variables, the most important conclusion one could draw from these models relates not to the nature of the achievement measures considered, but to the impact of the sample design on interpreting the explained variance, particularly in relation to the apparent size of the school social context effect.

6. CONCLUSION

6.1. Introduction

International interest in and concern with educational outcomes has never been higher. The globalisation of market economies and the belief that human capital is essential for economic competitiveness and success (and that it can be measured in an international assessment) has helped to fuel this interest (Kellaghan & Greaney, 2001; OECD, 1998). Widespread international interest in such outcomes is evident in the large increase in the number of countries participating in PISA since 2000 (32 in 2000, 41 in 2003, 55 in 2006).

The political priorities underlying surveys are key to understanding their purpose and design. Differences between IEA surveys and PISA can be traced back to the processes by which such surveys were established. The IEA is underpinned by theoretical concerns regarding the nature of education systems; the OECD established PISA with the express intent of gathering data pertinent to the competitiveness of knowledge-based economies. According to the OECD, the results of PISA are intended to provide a profile of knowledge and skills at or near the end of compulsory schooling; to yield information on the relative equity of education systems; and to indicate areas of education systems that could be improved. The monitoring of achievement over time is another important feature. PISA, for the first time, provides comparative data on educational outcomes for all OECD countries (and for an increasing number of 'partner' countries). Prior to PISA, data on outcomes were limited and not aligned to the political agenda of countries (OECD, 1992a, b).

In Ireland, PISA is used to monitor student achievements at post-primary level and is the only means of doing so (public examinations data are not suitable for this purpose since they are not standardised and designed to discriminate between individuals rather than describing achievements of the education system) (Greaney & Kellaghan, 1996; Kellaghan, 1995). Since December 2001, when the first results of PISA were published, there has been regular, and considerable, media commentary on the results for Ireland. Such commentary is, however, largely uncritical of PISA, focusing in a rather simplistic manner on country rankings and the distribution of students across proficiency levels,

and also including statements about the consequences of the results for the Irish economy. Government ministers have also referred to the results where PISA has been received uncritically, and the emphasis is on country rankings, proficiency levels, and the implications of the explanatory analyses for policy on educational disadvantage.

Despite this widespread interest, there has been no formal academic study of PISA in Ireland. Most of the critical commentary on the programme to date has come from the UK (e.g., Goldstein, 2004; Prais, 2003; Smithers, 2004), perhaps because of the public debate on the use of league tables as measures of school quality. The lack of critical commentary is surprising, particularly since the PISA tests represent a departure from previous surveys, where the express intent is *not* to measure what is taught and learned in school, but rather to assess the broader 'real-life literacy' skills of students. This puts an onus on countries to consider whether, in national context, outcomes as measured by PISA are desirable. Further, the utility of PISA results for curricular review and educational improvement is limited unless we have some information on how the PISA assessment and national curricula overlap and diverge. That sampled students tend to be dispersed across multiple education programmes and grade levels (second year to fifth year, or grades 8 to 11, in Ireland) makes it even more difficult to pin the results to a particular point in the education system. Further, there are a number of assumptions underpinning claims about the relative equity of education systems and explanatory analyses of the determinants of achievement which merit consideration in interpreting the results.

This thesis aimed to address these issues by providing an in-depth analysis of PISA to identify aspects of its design and how results are interpreted that may be problematic, and/or inconsistent with conclusions drawn by the media and government. Its focus was, therefore, more on the utility and interpretability of results than an in-depth treatment of the technical assumptions underlying such aspects as the test design and scaling (in any case, technical reviews of these aspects of surveys are already available; e.g., Blum, Goldstein, & Guérin-pace, 2001; Goldstein, 1995; Hambleton et al., 2005). Since reading was the major subject domain of 2000, and mathematics was the major focus of 2003, data used in the analyses in this dissertation on reading achievement are for the most part taken from PISA 2000, and data on achievement in mathematics are based on the PISA 2003 results. Three questions were addressed. First, what does PISA

tell us about the achievements of students in Ireland? Second, what does PISA tell us about the equity of achievement outcomes? Third, what does PISA tell us about the determinants of achievement? The three sections in this chapter which follow draw together the results of relevant analyses. The final section identifies some broader implications arising from the results, and suggests areas for further research.

6.2. What PISA Tells Us About Achievement

By international standards, the reading skills demonstrated by students in Ireland are high. The distribution of achievement also indicates that there are comparatively few low achievers (or, in OECD terms, fewer students are at risk of poor outcomes in occupational and social contexts in the future and by implication, there is increased likelihood of a more competitive knowledge-based economy). This pattern holds across the combined reading scale and the subscales, although performance on the Reflect subscale was slightly higher than on the others.

In mathematics, however, performance was lower in international terms. Irish students performed around the OECD average and exhibited an uneven profile of performance across the four mathematics subscales. While performance on the Quantity subscale was around the OECD average, performance on Uncertainty and Change & Relationships was significantly above the corresponding OECD averages, while performance on Space & Shape was considerably below the OECD average. Across all mathematics scales, broadly speaking, the same pattern of distribution is evident, whereby the performance of low achievers was comparatively high relative to low achievers internationally, while that of high achievers was comparatively low.

The fact that PISA is not intended to assess school-based curriculum, together with the fact that it is the only source of international data for monitoring educational outcomes at post-primary level in Ireland, highlight the relevance of analyses aimed at providing a better understanding of what the PISA results mean in Ireland. Both the PISA 2000 (Shiel et al., 2001) and PISA 2003 (Cosgrove et al., 2005) reports for Ireland have compared the Junior Certificate and PISA in terms of content and performance. Specifically, these entailed qualitative comparisons of the PISA assessment frameworks and Junior Certificate syllabuses, quantitative ratings of PISA test items in terms of the

Junior Certificate syllabus (the 'test-curriculum rating project'), and reports of statistical associations between achievement on PISA and the Junior Certificate.

In the case of PISA reading and Junior Certificate English, qualitative comparisons revealed differences in the formats of the tests (the Junior Certificate generally requires essay-type responses; PISA's response formats are considerably shorter); the type and length of the texts (the Junior Certificate entails studying lengthy literary texts, while almost all PISA reading texts are much shorter and there is a greater emphasis on functional texts); and the manner in which students' responses are marked (the Junior Certificate English marking schemes are more impressionistic and may allow credit for less precise answers than would be permitted by the PISA marking schemes). However, the reading processes/skills assessed are similar, particularly at higher and ordinary levels. The results of the test-curriculum rating project confirm these observations, whereby the processes underlying the majority of PISA reading items were rated as somewhat or very familiar. Therefore, it is not true to say that PISA reading is a curriculum-*free* assessment in the case of Ireland; the PISA measure of reading may be better described as one which is compatible with the national curriculum, if not providing a complete assessment of it.

In contrast, considerable differences between the style of the PISA 2003 mathematics test and Junior Certificate mathematics have been identified. These have been attributed to differences in the underlying philosophies of the assessments. PISA mathematics is rooted in Realistic Mathematics Education (e.g., deLange, 1998), emphasising horizontal mathematisation in concrete and authentic contexts. Junior Certificate mathematics is more formal and abstract (emphasising, particularly at higher and ordinary levels, theorems, proofs, and vertical mathematisation) (e.g., Oldham, 2002). The test-curriculum rating project indicated that most of the PISA mathematics items which assess concepts on the Junior Certificate mathematics syllabus are in areas such as applied arithmetic and measure. Many PISA mathematics items not on the syllabus examine concepts relating to probability (not studied until the Leaving Certificate) and spatial reasoning (not on either the Junior Certificate or Leaving Certificate). PISA mathematics does not assess any of the concepts covered in Junior Certificate trigonometry and geometry, and very little of algebra and functions. The manner in which Junior Certificate mathematics questions are marked allows more scope for merit

than the PISA marking schemes, since any reasonable attempt at a question on the Junior Certificate gets credit (in PISA, right-wrong marking is the method used to mark the vast majority of items) (Cosgrove et al., 2005). Students were expected to be moderately familiar with concepts underlying the PISA test items, but generally unfamiliar with the contexts in which these concepts were embedded, which is consistent in the observed differences in the underlying philosophies of the two assessments. A backwards-mapping of Junior Certificate mathematics items onto the PISA mathematics framework (Close & Oldham, 2005), which complements the test-curriculum rating project for mathematics, also indicates that Junior Certificate test items are in abstract, intra-mathematical contexts to a much greater degree than the PISA test items. Further, although it was intended that some of the Junior Certificate Examination questions would assess application of mathematical concepts in somewhat novel contexts, in practice, this is not the case. The vast majority of Junior Certificate questions at all syllabus levels were classified by Close and Oldham (2005) as belonging within the Reproduction competency cluster. This suggests that students can pass the Junior Certificate mathematics examination largely through the mechanical reproduction of learned routines, an observation which is present in the Chief Examiners' reports on Junior Certificate mathematics, and also noted in the NCCA review of mathematics education (NCCA, 2005).

When test-curriculum project ratings were aggregated to the level of the student and associations with achievement on the corresponding PISA domain computed, results suggest that familiarity with mathematics concept is most strongly predictive of success on PISA mathematics, while familiarity with reading process and context of application had moderate associations with achievement on PISA reading.

Associations between performance on PISA and the corresponding Junior Certificate subject indicate moderate overlap in achievement, with correlations of around .70 for both reading and mathematics. However it was noted that correlations between the PISA domains of mathematics and reading, and between Junior Certificate mathematics and English, are similar, at .70. It was suggested in Chapter 2 that other factors, such as the manner in which results are marked, and differences in the stakes associated with the tests, should be considered in interpreting these associations. It was also suggested

in Chapter 2 that an overall measure of association might disguise differences at the extremes of the achievement distribution.

Chapters 1 and 2 also identified some limitations of the test-curriculum rating project analyses. Perhaps most importantly, the results cannot be interpreted in an international comparative context. Only eight or so countries appear to have analysed PISA with respect to their curricula, and the analytic frameworks for doing so are not comparable. This imposes limitations on the utility of the PISA results, particularly in the area of mathematics, where countries must somehow identify areas for educational improvement without a clear description of how curricula in other countries compare with theirs vis á vis what PISA assesses. It was noted in Chapter 2 that the response of the PISA Governing Board and the OECD Secretariat to a request for a curriculum analysis project by the UK DfES (2004) (with participation on a voluntary basis) was not entirely satisfactory. [One imagines that when the results for PISA 2006 are published, where science is the major domain, that requests similar to those made by the DfES (2004) for comparative analyses of science curricula will be made.] A resolution to this issue at OECD level would be highly desirable, both to offer the possibility of enhancing the utility of PISA in countries where curricular review and reform are policy priorities, and also to enhance the credibility of the PISA survey itself, particularly when comparisons with TIMSS are planned as one of the themes for secondary analyses of the PISA 2003 data (OECD Secretariat, 2005, February).

A second limitation of the results of the test-curriculum rating project is the poor explanatory power of the results. It appears that, once performance on PISA is adjusted for student ability (the adjustment, comprising performance on the Junior Certificate, is not ideal but the best available), the relationship between performance and expected familiarity is not statistically significant. Further, the relative familiarity of items corresponding to the various subscales on both reading and mathematics do not map onto the observed differences in the mean performance on the PISA subscales. For example, one would expect, based on the Irish mean on the Uncertainty subscale, that the items on this scale would be accorded a relatively high familiarity rating, while the low mean score on the Space & Shape subscale suggests that items on this scale would receive a low familiarity rating, but this is not the case. Part of the problem relates to the lack of data from other countries with which to compare these ratings. It may also be the

case that the item characteristics used as a basis on which to make ratings were not the most relevant or appropriate. Nohara's (2001) comparative analysis of the PISA, TIMSS and NAEP items included a consideration of other aspects of the test items such as the presence of cross-curricular elements and amount of steps required in the reasoning process, and these and other characteristics, particularly the marking schemes and how they are applied, may have provided relevant information in the case of PISA and the Junior Certificate. Another possibility is that the ratings were made at the wrong level of specificity and/or need to be supplemented with analyses of the curriculum as implemented. Analyses of the PISA 2003 mathematics teacher questionnaire (Cosgrove, Shiel, Oldham, & Sofroniou, in preparation), which asked teachers to rate the relative emphasis placed on various aspects of the PISA subscales as described in the PISA mathematics framework, provide results which are more consistent with the pattern of performance observed. For example, the reported emphasis placed on five key concepts and skills associated with the Space & Shape subscale (as defined in the PISA assessment framework) was relatively low.

Other limitations of the test-curriculum rating analyses relate to the design of PISA itself. First, PISA's design is cross-sectional so there is no opportunity to measure achievement gains across time and relate these to curricular exposure; a longitudinal design would be more appropriate for explanatory analyses of opportunity to learn (Goldstein, 1995; 2004). Second, the sample design does not permit achievements to be linked to class-level or teacher variables, so links to student achievement are limited to the intended curriculum rather than an analysis of the implemented curriculum. There is a need to supplement such ratings with classroom-based research, and the lack of classroom-based research with respect to mathematics teaching in Ireland has been noted (Lyons et al., 2003). Recent observational research on the topic suggests that didactic teaching methods which give little scope for conceptual development and emphasise rote learning are rife in post-primary mathematics classrooms in Ireland (Lyons et al., 2003). However the research has been confined to just 10 schools so the generalisability of the results is limited.

The literature review also suggested that there are considerable disparities in the achievements of students at higher, ordinary and foundation levels in both reading and mathematics (Cosgrove et al., 2005; Shiel et al., 2001). A comparison of the

percentages of students at each syllabus level who score at each PISA proficiency level also demonstrates the large disparity in achievement between the syllabus levels with 90% of students taking foundation-level English scoring at or below Level 1 and about 28% of ordinary-level students were at or below Level 1. In contrast, just 2% of higher-level students were at or below Level 1. At the other extreme, 10% of ordinary-level students, and 55% of higher-level students were at Levels 4 or 5 on PISA reading. In mathematics, the achievements of close to three-quarters of foundation-level students were at or below Level 1. At ordinary level, 22% of students were at or below Level 1, and around one in eight at Levels 4, 5 or 6. Under 2% of higher-level students were at or below Level 1, and three-fifths at Levels 4, 5 or 6.

If we accept that PISA is an appropriate indicator of performance, the substantial percentages of ordinary-level students scoring at or below Level 1 in both reading and mathematics suggests that the standards being applied at this level, which is geared at average-ability students, may be too lenient. The findings also suggest that many students taking English and mathematics at both ordinary and foundation levels lack basic reading and mathematical skills and, if the OECD's interpretation of the likely consequences of achieving at Level 1 and below is correct (OECD, 2001b; 2004c), then these students are unlikely to achieve their potential in the future. On the other hand, a substantial minority of ordinary-level students are achieving at the more advanced levels of the PISA proficiency scales, which suggests that they are not being sufficiently stretched by the material they are studying in school and possibly, that they would be better suited to higher-level courses. This is particularly evident in mathematics, where the comparatively poor performance of higher achievers in Ireland on PISA 2003 has been noted.

A comparison of the percentages of students at each syllabus level attaining grades A-F on the Junior Certificate with the percentages of students at each PISA proficiency level indicates that the 'fail' rate (below a grade D) is low across all syllabus levels; the large percentages of students at or below Level 1 particularly at ordinary and foundation levels would predict a higher 'fail' rate. As noted already, the leniency in some aspects of the marking of Junior Certificate mathematics and English may be allowing some weak students to attain a grade C or D on the Junior Certificate, while the more

dichotomous and precise criteria applied to the marking of the PISA items result in zero credit for some test items attempted by these students.

A comparison of the percentages of students taking Junior Certificate English and mathematics at each syllabus level awarded each letter grade suggests that the grades assigned to students at higher level in both subjects may be amenable to interpretation in terms of the PISA proficiency levels. Aligning the letter grades applied at ordinary and foundation levels to the PISA proficiency scales is more problematic and is likely to be related to the fact that the public examinations are designed to discriminate between higher achievers more so than those at the average or lower points of the ability distribution.

The results of a series of analyses reported in Chapter 4 which were designed to identify an alternative way in which to scale the Junior Certificate English and mathematics results based on achievement on PISA provide evidence that discrimination at the lower level of *both* of the achievement distributions is poor. In fact, in both English and mathematics, the average achievements of students on PISA attaining below a grade D at ordinary level, and below a grade A at foundation level, could not be distinguished from one another. Evidence of stretching at the upper end of Junior Certificate mathematics was found; this was not the case for English. This suggests that the discrimination between an A and a B grade at higher level in mathematics is greater than for English. The inability of the PISA achievement measure to distinguish between students achieving grades E or F at ordinary level, and below grade A at foundation level, should be interpreted with respect to differential response rates of low and average to high achievers (discussed in more detail later in this section), resulting in lower numbers at the lower end of the Junior Certificate Performance Scales (JCPS) than may have been observed in the population. It may be the case that, had more ordinary- and foundation-level students participated in PISA, the lower points of the JCPS might have been empirically distinct. Moreover, the low numbers of test items with difficulty levels at or below Level 1 do not allow for a meaningful distinction between those students at Level 1 and those below Level 1. This distinction between at Level 1 and below it merits revisiting since the combining of students at and below Level 1 may not be warranted, particularly in use of targeted interventions of the lowest achievers. In any case, the wider implication is that the PISA achievement measures of

both reading and mathematics are of very limited use in describing the achievements of foundation-level students and students attaining at the lower end of the letter grades at ordinary level since most of these students are 'off the scale'.

The choice of the 'preferred' JCPS was validated though a comparison of the results for 2003 (in the case of reading) and 2000 (in the case of mathematics). Broadly speaking, these provide support for the choice of the scales, but they also suggest that the lower ends of the scales are prone to fluctuations in both subject areas, and in mathematics, fluctuations were also observed at the upper end. The reasons for this are difficult to ascertain. They may relate to the fact that there were more PISA items with difficulties corresponding to the extremes of the ability distributions in the particular PISA cycle corresponding to when a particular domain was the 'main' domain; with the decreased reliability associated with the grades assigned to the lower end of the ability distribution of the Junior Certificate; to changes in the population of students taking the tests; or changes in the difficulty or other aspects of the Junior Certificate Examinations. Perhaps the measurement of achievement in students of lower ability is in essence less reliable than measures of achievement of students at medium and high levels of ability. Wiliam (2004) comments, with respect to the reliability of educational assessments, that

> Even where the test does sample the whole breadth of the domain, it is often the case that the sample is spread so thinly that the result is much more a consequence of the particular items selected for the test than any indication of the candidate's capability in the domain of interest. (p. 3)

If this is true, then estimates of the performance of low and high achievers will be particularly prone to item selection effects, since it has already been shown that the PISA tests contain relatively few items at the extremes of the ability distribution.

The distribution of performance on the 'preferred' scale for mathematics was similar across the four mathematics subscales; it was hoped that some fluctuations may have been observed which might in turn have lent themselves to further insights into the knowledge and skills of Irish students which might have been related to the mathematics curriculum, but this did not turn out to be the case. In the case of reading, relatively strong performance of low-achieving students on the Reflect subscale was evident, although the reasons for this are unclear. One can speculate that these items

may have resulted in higher interest and engagement on the part of these students, and/or that the item formats (proportionately more of which involved written responses) may have been more familiar to such students.

Addressing the question as to what PISA can tell us about the achievements of Irish students also entailed a consideration of potential sources of bias in the achievement estimates. The review of PISA's sampling standards (e.g., Adams & Wu, 2002) suggests that in general, PISA is in line with, or exceeds, standards associated with population coverage and response rates. However, more recently, several authors (e.g., Beaton et al., 1999; DfES, 2005; Monseur & Wu, 2002) have called attention to the problems associated with the methods used to adjust for student and school non-response, and it has been pointed out that there are no agreed standards in the quantification of bias arising from non-response.

In PISA, the school-level adjustments take account of the explicit and implicit stratifying variables, thereby reducing or eliminating non-response bias, but the student-level adjustments assume that participants are the same as non-participants, which is unlikely to be the case. Further, recent research using simulated data (Monseur & Wu, 2002) has shown that the efficiency of non-response adjustments is related to the manner in which achievement variance is partitioned between and within schools. In countries with small between-school variance (including Ireland), the non-response adjustment at the school level is efficient at eliminating bias in the achievement estimates, whereas those at the student level are inefficient. Bias is even more likely to arise when propensity to participate is related to achievement at the school level, which was found to be the case in Ireland in both 2000 and 2003. A third reason to be concerned about non-response bias in Ireland is that the student-level response rates in both 2000 and 2003, although meeting the agreed-on minimal standards, are on the low side in comparison with many of the other participating countries. The possibility of non-response bias does not only apply to Ireland. For example, a correlation of .20 or higher was found between school-level achievement and propensity to participate in 15 of the 32 countries in PISA 2000, and in seven of these, the correlation exceeded .30 (Monseur & Wu, 2002).

Analyses in Chapter 3 explored, using multiple logistic regression, whether non-responding schools and students in PISA 2000 and PISA 2003 were comparable in a number of respects to those who did respond to the assessments. Schools were compared by type (sector), sex composition, designated disadvantaged status, size, and school mean performance. There was no evidence of bias in the 2000 sample. In 2003, secondary schools were under-represented, but the non-response adjustments take school type into account so the 2003 school sample may also be regarded as free from non-response bias, particularly when participating and non-participating schools did not differ in average performance on the Junior Certificate in either 2000 or 2003.

The Junior Certificate achievements of absent and refusing students in both PISA 2000 and 2003 were significantly lower than students who participated. A comparison of participating and non-participating students was also made by sex, grade level, syllabus level, school designated disadvantaged status, school sex composition, and school type (sector). Using multiple logistic regression, it was found, in PISA 2000, that second and fourth years were less likely to participate than third years; that ordinary and foundation-level participation rates were significantly lower than higher level; and that students in vocational schools were less likely than students in both secondary and community/comprehensive schools to participate.

In 2003, fourth and fifth year students were significantly less likely to participate than third years, whose participation rate did not differ to second years. Differences between the participation rates of higher, ordinary and foundation levels were highly significant. Students taking Junior Certificate mathematics at higher level were most likely to participate. In PISA 2003, students in community/comprehensive schools were significantly less likely than students attending secondary and vocational schools to participate. Two additional variables relating to school factors were in the model of participation for 2003 – designated disadvantaged status (students in designated schools were less likely to participate) and sex composition (students in single-sex schools were more likely to participate).

These analyses suggest differential non-response in 2000 and 2003 and this may be related to the fact that the major domain of the assessment was different in the two years. They are strongly indicative of an upward bias in the overall mean scores of

students in both 2000 and 2003. Therefore, conclusions drawn about reading and mathematics standards of lower-achieving students are particularly problematic since many students in this group were not even present for the PISA assessment, even though they were eligible to participate. The results also suggest that conclusions about some subgroups (e.g., students in vocational schools in 2000 and in community/comprehensive schools in 2003) are not as reliable as those for other subgroups (e.g., students taking English and mathematics at higher level for the Junior Certificate).

Given the emphasis in national and international reports, in the Irish media, and by Irish government ministers on the percentage of students at or below Level 1 as an indication of the extent to which a population has low levels of literacy, and the finding from the re-analysis of the IALS data (Coulombe et al., 2004) that a reduction in low levels of literacy rather than an increase in high levels of literacy is associated with economic growth, a review of the procedures to adjust for non-response to obtain a more accurate estimate of the percentage of students with low levels of literacy is warranted. This is particularly worthy of consideration when one also considers that the efficiency of non-response adjustments vary according to how achievement variance is partitioned between and within schools and according to the strength of the relationship between propensity to participate and achievement (Monseur & Wu, 2002). Countries in PISA vary widely on how much achievement between schools varies, as well as in the relationship between achievement and propensity for participation.

The main conclusion that one might draw from these analyses is that, in the absence of corrective measures for this bias and with few test items at the extremes of the achievement distribution, it makes little sense to use the PISA data to develop educational policy on issues related to low achievement in either reading or mathematics.

6.3. What PISA Tells Us About Equity in Achievement Outcomes

A second theme examined in this thesis considers claims about the relative equity of education systems, particularly with reference to the sample design. The percentage of total achievement variance which is between schools is interpreted as a measure of homogeneity of schools in an education system (Postlethwaite, 1995) and the OECD

(e.g., 2001a; 2004c) has taken the interpretation further, taking low between-school variance as an indication of the educational equity of a system. This notion of equity has also appeared in Irish media reports on PISA and in ministerial speeches. This simplistic interpretation of equity is problematic since the choice of sample design affects the manner in which achievement variance is partitioned between and within schools.

Analyses in Chapter 5 compared the variance components associated with mathematics achievement for TIMSS 1995 (which used intact-class sampling and reported the between-school variance statistic on the basis of one sampled class per school) and PISA 2000. These showed that, for some countries, the between-school variance statistic is similar in both studies; for others, including Ireland, between-'school' variance is much higher for TIMSS than for PISA. This is consistent with other studies reviewed in Chapter 2 which suggested that, in Ireland, partitioning students on the basis of ability occurs within schools, perhaps more so than between them. A re-analysis of the variance components of the TIMSS data for the Irish participants in Chapter 5 confirmed this. The net consequences of these analyses are first, that one is likely to draw very different conclusions about the relative equity of education systems depending on whether the sample design is age-based or grade-based, particularly if class allocation is made on the basis of ability, and second, in the case of PISA, the low between-school variance disguises large achievement differences between classes. Unfortunately, the available data do not allow for robust inferences since they entail comparisons across different surveys, such that many differences (e.g., test content, population surveyed) have not been controlled for. Nonetheless, they should be taken as *prima facie* evidence of the relevance of the issues. There is a paucity of research which examines the achievement variance of students in Ireland using a three-level model, whereby within-school variance is partitioned into between-class and within-class components. There is also a lack of available data from surveys which used a combined age-/grade-based sample design with which to further explore this issue.

The analyses reviewed in Chapter 2 also indicated that, not only does the portioning of achievement variance depend on the sample design, it may also depend on both the subject area considered and whether it is intended to be curriculum-neutral or not, with the highest between-school variance associated with school-dependent, curriculum-

sensitive measures (although much of the relevant research in Ireland was conducted during the 1970s and one cannot say whether or to what extent the findings still apply). Given that reading is a more school-independent, generic skill, it is hardly surprising that the variance components for PISA reading and Junior Certificate English were similar to one another. In the case of mathematics, consistent with previous research, between-school variance was somewhat higher for the Junior Certificate compared with PISA; however, it is possible that the observed differences would have been much greater, had the PISA sample design employed the selection of intact classes. A comparison of these results with the variance components for TIMSS mathematics and the 1996 mathematics examination suggests that this is the case. Again, unfortunately, one cannot be confident about these inferences since they entail a comparison of two different surveys. These comparisons do indicate, nonetheless, that it would be worthwhile exploring the dual themes of sample design and test content further within a single survey dataset, if one were available.

Finally, the analyses of non-response bias which compared (for both 2000 and 2003) total and between-school variance for all available Junior Certificate achievement data with achievement data for only those students who participated in PISA suggest that, in the case of Ireland at least, the between-school variance statistic appears to be unaffected by student non-response (where a downward bias in between-school variance was expected). However, between-student variance may have been underestimated (as predicted), which suggests that, had all students participated, the standard deviations associated with achievements on PISA might have been bigger, which in turn would call into question claims that overall performance is fairly homogenous.

The principle conclusions from these analyses is that the sample design appears to exert considerable influence on estimates of between-school variance; that the nature of the achievement measure used is of potential relevance in considering the meaning of between-school achievement differences; and that student non-response may create a downward bias in total achievement variation.

6.4. What PISA Tells Us About the Determinants of Achievement

Widespread use of multilevel explanatory models in the international and national survey reports, which have tended to emphasise the relative impact of social intake and

school/class variables on achievement, are underpinned by a number of assumptions. The results of these have been received largely uncritically, and there is evidence that the PISA analyses are being used to inform policy on educational disadvantage in Ireland. Therefore, the third theme of the thesis considered what PISA tells us about the determinants of achievement and whether conclusions drawn are problematic.

Several general limitations of these models were noted in Chapter 2. These include the lack in most of the models of a measure of student intake, thereby inflating the apparent effect of school and student social background on achievement; general difficulties with making causal inferences from the models; and the lack of adequate measures of school and class inputs and processes (Goldstein, 1997; Raudenbush & Willms, 1995).

The OECD has reported results of multilevel models for both PISA 2000 and PISA 2003 (e.g., OECD, 2001b; 2004c; 2005c) with a particular focus on the impact of school and student SES on achievement. A recent report on school effects which used data from PISA 2000 (OECD, 2005c) indicated that variables relating to school and class inputs and processes have little additional explanatory power over and above SES, and that the explanatory power of these was lower in Ireland compared with the OECD average. These analyses were identified as being problematic in Chapter 1 for a number of reasons. First, the OECD fails to acknowledge that the sample design may affect the extent to which school and student SES are related to achievement. It has already been shown that the manner in which achievement variance is partitioned between schools can vary depending on whether intact classes or a random within-school sample is selected, which in turn can influence the available variance to be explained between and within schools. Second, it also fails to acknowledge that the extent to which variables related to school inputs and processes may vary depending on whether the achievement measure is intended to be curriculum-sensitive or not, as well as whether the measure is school-dependent (e.g., of mathematics, science) or not (e.g., of a more generic skill such as reading). A review of studies which included explanatory analyses of achievement of students in Ireland in Chapter 2 suggested that all of these factors should be considered when interpreting the results of explanatory analyses (although the most relevant of these to the particular issues under consideration are close to 30 years old). Third, the multilevel models reported by the OECD do not include some variables

which are relevant to the structure of the education system in Ireland, such as school sector.

Although achievements on PISA and the Junior Certificate have been compared within a multilevel modelling framework in the case of PISA 2000 (Sofroniou et al., 2000; 2002), the models do not address the issues raised here. First, the relative contributions of student and school SES are not considered separately. Second, there is a paucity of school-level variables included in the models. Third, the models include students dispersed across multiple grade levels. Differences in the difficulty and/or marking of the Junior Certificate Examinations across the years considered complicate the interpretation of results, especially if conclusions are to be drawn about the Irish education system at a particular point in the system (i.e., at the end of Junior Cycle).

The multilevel analyses presented in Chapter 5 attempted to address these issues and limitations. On the basis of the literature review, several hypotheses and research questions were identified. Six multilevel models (PISA 2000 reading, 2000 Junior Certificate English, PISA 2003 mathematics, 2003 Junior Certificate mathematics, TIMSS 1995 mathematics, 1996 Junior Certificate mathematics), which investigated the contributions of a common set of school-level and student-level variables, were compared. As with comparisons of the variance components reviewed in the previous section, one should be cautious about comparing the results of multilevel models which were constructed using data from different surveys; nonetheless, the analyses use the best available data and every attempt has been made to maximise the comparability of the models.

The first hypothesis investigated was that mathematics achievement would be more sensitive to school-level effects than measures of English/reading. This received some support. Of nine school-level variables other than SES examined in the models, just one (disciplinary climate) was in the final model of 2000 Junior Certificate English and PISA 2000 reading. The two models of Junior Certificate mathematics retained additional school-level variables, particularly the model for 2003 (which included school type, disciplinary climate, and building quality). However, the model of PISA mathematics did not require any variables at the school level other than school SES, but

this finding supports the argument that PISA is the least curriculum-sensitive of the mathematics measures considered.

The second hypothesis was that Junior Certificate mathematics would be more sensitive than both PISA mathematics and TIMSS mathematics to school-level effects if it is more curriculum-sensitive. This hypothesis also received support. The final model for Junior Certificate mathematics in 2003 had three significant school-level variables other than SES. The school-level variables (other than SES) in the model for 1996 Junior Certificate mathematics also explained proportionately more of the within-school and between-school variance compared with TIMSS 1995.

It was also hypothesised that, because of similarities in the reading processes assessed in PISA reading and Junior Certificate English as well as the less school-dependent nature of reading, that the models for these would be similar. Broadly speaking, this received support although minor differences were found (e.g., the slope associated for gender varied significantly across schools in the model for Junior Certificate English but was constant in the model for PISA 2000 reading).

Regarding the impact of social intake, it was hypothesised that the association between school-level SES and achievement would be strong in all models. Strong support for an effect associated with social intake was found, and its relationship with achievement was curvilinear rather than linear in five of the six models, with a tapering off of the strength of the effect at higher values of school SES. This is at odds with other explanatory models of Irish student achievement reviewed in Chapter 2. However, two additional models of achievement in 2003 (Junior Certificate mathematics and PISA mathematics) which used Junior Certificate fee waiver rather than combined economic, social and cultural status as the school social context measure had a linear association with achievement. This suggests that income-related measures may be more likely to have a linear association with achievement than composite measures.

It was also hypothesised that the social context effect would be weaker in the models of mathematics compared with English/reading (given that reading achievement may be more strongly associated with social background). This hypothesis did not receive support, regardless of whether one considers achievement on PISA or the Junior

Certificate. This contrasts with Sofroniou et al.'s results and again suggests that the manner in which social context is measured (i.e., whether income-related or a composite) may be relevant in considering this issue.

It was further hypothesised that, if students are clustered within classrooms on the basis of social background (as well as ability), the strength of the social context effect for TIMSS 1995 and 1996 Junior Certificate mathematics would be stronger than the social context effect associated with models for 2000 and 2003 since the former are clustered on the basis of class membership; the latter are distributed across classes. This hypothesis received strong support and indicates that estimates of impact of social intake should take cognizance of the sample design. However, there is no difference between the PISA and TIMSS samples with respect to the extent of social segregation within and between classes/schools (as indicated by the proportion of variance in SES between classes/schools). Therefore, the differences could be due to clustering of students within classes based on factors which somehow mediate the relationship with SES and achievement, or the fact that different, although comparable, components, made up the SES measure used in the models of TIMSS and PISA.

The existence of a cross-level interaction between the social intake and gender whereby the slope would be steeper for males was expected in at least some of the models. However, no such cross-level interactions were found. It is possible that the manner in which SES is measured at *both* school and student levels is also of relevance, whereby an income-related measure at both levels may have produced a cross-level interaction. The lack of alternate student-level SES measures based on income in the PISA and TIMSS datasets prevent this issue from being explored in more depth in the context of the present research.

Overall, results support the argument that the interpretation of multilevel models should take account of the nature of the achievement measure used and perhaps even more so the nature of the sample design, particularly in examining the size of the social context effect and the additional variance explained by school-level variables (other than SES). They also point to the need for more research on the nature of the impact of social intake (e.g., why it is linear rather than curvilinear in some models, and why, in some models, cross-level interactions are detected, while in others, they are not).

6.5. Concluding Remarks

This section considers several areas which merit further research in national and international contexts; some aspects of PISA's survey design which could be modified to enhance the interpretation of findings; and how findings relate to national policy concerns. The literature review and results of analyses suggest 13 key implications.

1. The results for Irish students on PISA 2000 reading have made a relatively small impact on educational policy, since they tend to affirm that Ireland is successful in providing an education which has produced a relatively skilled student population of readers. However, the outcomes on mathematics have resulted in more controversial media commentary, including a call to review mathematics education in Ireland generally. The NCCA (2005) has recently undertaken a review of mathematics education at post-primary level, and the influence of the PISA survey is clearly evident in this review. One might argue, therefore, that PISA has an 'enlightenment' function with respect to mathematics education, despite, or perhaps even because, there is a mismatch between what PISA mathematics and Junior Certificate mathematics assess. The NCCA review takes a broad view of the issues. In describing the context of the review, it is stated:

   > The review is not simply an exercise in syllabus revision – although this may be an outcome of the review – but rather a more fundamental evaluation of *the appropriateness of the mathematics that students engage with in school and its relevance to their needs*. It must take into consideration broader reviews that are taking place (the implementation of the primary school curriculum; junior cycle) and the proposals being developed for senior cycle education. (NCCA, 2005, p. 3, italics added)

   The PISA assessment framework was developed by international experts in collaboration with participating countries and emphasises the importance mathematical skills in the context of globalised economies. Given, too, that PISA purports to assess mathematics skills that are *relevant* for current and future participation in wider society, this suggests that the NCCA should pay particularly close attention to what PISA mathematics assesses, whether this is relevant to students in Ireland, and what implications this has for the teaching and learning of mathematics. It suggests in turn a fundamental consideration of

whether an approach to teaching and learning mathematics that is based more in Realistic Mathematics Education (e.g., de Lange, 1998) is desirable in the Irish context. If so, this implies a rebalancing of the curriculum to focus less on abstract (and at times mechanistic) approaches to mathematics.

2. The very poor performance of a substantial minority of ordinary-level students in both reading and mathematics is evident in the PISA data. In contrast, some ordinary-level students achieve at the high end of the PISA proficiency scales. However, as already pointed out, little is known about how and why students come to take the syllabus levels they do. The lack of research on syllabus take-up was noted in Chapter 2. The only available research has been commissioned by the NCCA as part of a review of the Junior Cycle. This is longitudinal in design and gathers both qualitative and quantitative data from approximately 900 students in 12 schools (Smyth et al., in preparation; at the time of analyses, students were in second year). Preliminary results indicate that take-up is affected by the social intake of the schools over and above student ability, and that the effects of the school's social intake may be magnified though differences in teachers' expectations. In the first wave of this longitudinal research, Smyth, McCoy and Darmody (2004) found that while 94% of the schools use the results of entrance examinations to group students (usually in English, Irish and mathematics), the methods used to assess the students were extremely varied (indeed in 42% of cases, tests were designed in-house). These findings should be considered in the absence of concrete guidelines on assigning students to syllabus levels. The PISA-Junior Certificate analyses reported in Chapters 2 and 4 suggest that the mismatch is quite marked at ordinary level. In its review on mathematics education at post-primary level, the NCCA (2005) cites timetabling and class allocation as additional relevant factors and stress the importance of exposing the maximum number of students possible to higher-level mathematics. Syllabus allocation raises broader concerns about equity in educational opportunities. Allocation at Junior Certificate level plays a major role in 'locking' students into a syllabus level for the Leaving Certificate, which in turn dictates students' chances of accessing various third-level and post-secondary education (e.g., Millar & Kelly, 1999). According to Smyth (personal communication, October 4, 2005), "the key issues to be addressed in policy

terms are probably expectations (both teacher and student) and timing of allocation to subject level". The findings of Smyth et al. cannot, unfortunately, be generalised to the wider student population due to the small sample size. Therefore, research on the reasons for syllabus take-up is merited in the context of a larger-scale survey. Factors relating to class allocation, timetabling, teacher characteristics and expectations, student ability, attitudes and expectations, peer effects, and parental involvement and expectations should all be considered.

3. PISA and the Junior Certificate fulfil different purposes. Nonetheless, the percentages of ordinary- and foundation-level students who receive a grade E or F on the examinations is lower than the distribution of students on the PISA proficiency levels might have predicted, particularly at ordinary level, where the courses are geared towards average-ability students. As the marking schemes for PISA generally allowed less opportunity for merit for poor responses than the marking schemes for the Junior Certificate, this raises a question about whether there is too much room, at present, to 'pass' students who appear to have serious literacy and numeracy problems. These pass rates may serve to disguise the numbers of students entering Senior Cycle who might benefit from additional support with learning. Alternatively, or additionally, a proportion of the open-ended items in Junior Certificate Examinations in both of these subject areas could be replaced by a series of multiple-choice questions. The marking of these would be more efficient and results more reliable, and students' responses on the multiple-choice portion could be used by markers in instances where there was some degree of interpretation as to the relative merit of a student's response to an open-ended question.

4. The concerns raised by Close and Oldham (2005) that making the marking schemes available to the general public may result in marks-based instruction have not been verified and perhaps are not possible to verify. In a more general sense, it would be desirable to supplement students' responses to questions on the Junior Certificate Examinations with graded samples of their work over the course of the Junior Cycle to enhance the validity of the assessment as well as its reliability (noted in point 3). In its update on the Junior Cycle Review (NCCA, 2004), the NCCA has noted the mismatch between the Junior Certificate

programme's objectives of breadth and balance and its mode of assessment. However, at present, there do not appear to be any proposals to change the mode of assessment for certification at the end of the Junior Cycle.

5. The NCCA might consider commissioning a review of the content of Transition Year mathematics modules, since little is known about the content of these. It may be the case that Junior Cycle may not be the optimal (or even appropriate) point in the system at which to develop the types of mathematical problem-solving suggested by PISA. There is some anecdotal evidence which suggests that a real-life, cross-curricular approach to mathematics in Transition Year enhances students engagement in, and confidence with, mathematics (McCloughlin, 2005). It may be possible to capitalise more on the educational opportunities presented in Transition Year to enhance students' mathematical skills in real-life contexts.

6. The large differences in the size of the achievement variance between 'schools' in PISA and TIMSS suggest that in Ireland at least, careful account should be taken of the sample design in interpreting the results. The results of the three-level models of TIMSS students, which showed that between-'school' variance is lower when two intact classes rather than one are included, calls into question the appropriateness of using single intact-class samples for drawing conclusions about between-school variance in Ireland generally. It would seem important to communicate this message clearly in both national and international reports, particularly when comparisons across surveys are likely (see O'Leary, 2001).

7. At present, there does not appear to be available dataset which has incorporated both an age-based and a grade-based design. This is unfortunate since it would allow more robust inferences to be drawn about the consequences of both sampling approaches for the interpretation of variance components in Ireland; better inform policy on how achievement differences between schools and classes operates; and add insights into the appropriateness of various sample designs for addressing various policy and research questions in Ireland. There may be merit in supplementing the PISA sample design in Ireland in future surveys with the selection of intact classes at a particular grade level. Of course,

careful consideration would need to given to the target grade level, target age group and achievement measures to be used. Alternatively, a national survey which includes both sample approaches could be implemented at post-primary level. Indeed, if the post-primary schools database of the Department of Education and Science included data on the class membership of Junior Certificate and Leaving Certificate students for the core subjects, then that database could provide a potentially powerful tool for analyses of achievement variance.

8. The analyses presented in Chapter 5 indicate that there appear to be potentially significant consequences arising from the choice of SES measure in explanatory analyses of achievement. Income-based measures suggest a uniform increment in the relative advantages accrued, while composite measures suggest a ceiling effect. Furthermore, the relative disadvantage of subgroups of the population would appear to vary depending on the SES measure used. Not only this, but the relative impact of social background may be much larger in analyses of surveys whose results are based on a sample of one intact class per school. However, these assertions are at the level of conjecture since they have not been systematically investigated within a single survey dataset. Further research into identifying which types of SES measures are associated with cross-level interactions and linear/curvilinear effects and why their effects appear larger in class-based samples is therefore warranted on the basis of findings arising from Chapter 5. This research could begin by reviewing available datasets, both nationally and cross-nationally, to assess whether they are suitable for addressing these issues.

9. In a more general sense, however, only so much information can be gleaned from a multilevel modelling framework. As noted earlier, multilevel models are limited in allowing causal inferences and cannot adequately capture contextual factors and processes at play, such as how school culture operates, how teacher expectations manifest themselves, or why, in some schools, peer culture values high achievers, while in others, high achievement is frowned upon. Therefore, future analyses which aim to enhance our understanding of the social context effect and how and why it operates need to include not only empirical,

quantitative analyses (perhaps as an initial step), but also qualitative observational research (see Thrupp, 1999).

10. The manner in which the OECD and PGB have addressed the issue of curricular variation in countries participating in PISA has not been entirely satisfactory. It should be revisited. There may be merit, for example, in developing a means for smaller-scale, optional projects on areas such as curriculum where subsets of countries with a policy interest in curriculum review and improvement can work together without adversely affecting the main PISA activities. In this respect, it is promising that the terms of reference for PISA 2009 appear to offer some scope for flexibility, incorporating a 'modular' approach with a 'core' PISA design and three optional components (although curriculum does, as of yet, not feature in these modules) (OECD Secretariat, 2005, September).

11. There is also a need to revisit the methods used to adjust for non-response. Understandably, PISA is constrained in terms of the extent to which its design may be changed, particularly since one of its purposes is the monitoring of achievement trends. Therefore, if this aspect of the design cannot be changed, the OECD needs to be more transparent about non-response bias, and develop agreed-on standards that take account of country differences in between-school variance and of the relationship between achievement and propensity to participate. The OECD should also offer countries that wish to obtain unbiased achievement measures (e.g., for a more precise indication of the percentage of students who might be described as low achievers) technical advice on how results might be adjusted for this purpose.

12. In Ireland and other participating countries with relatively low student participation rates and a known relationship between achievement and propensity to participate, further efforts need to be made to increase the participation rates of students. The results in Chapter 3 identify subgroups of students who are less likely to participate in Ireland and therefore might fruitfully be used to develop strategies on increasing participation in future survey cycles. In a broader sense, the reasons why such students do not attend such assessments needs to be examined since at present, it is not known whether

this is arising from disengagement on the part of the students, advice from parents or school staff, or a complex mix of these and other factors.

13. The low number of items at the extremes of the ability distributions on PISA was noted. However, what is not known is whether including more items with difficulties at the extremes of the distribution might increase the reliability of estimates of the percentages of students at or below Level 1, and at the highest proficiency levels. In particular, it would be of interest to see how the addition of items to the low end add to our understanding of the skills (and, possibly, needs) of the lowest achievers. Therefore, consideration should be given in any technical review of PISA (such as that carried out by Hambleton et al., 2005) to conducting analyses, using simulated data, which investigate the effects of including a higher number of test items which assess the skills of very high and very low achievers.

To summarise, the 13 key points in this section suggest areas for future research which relate to practical and substantive issues as well as to more theoretical and technical ones. Points 1 to 5 would appear to relate more to practical concerns; namely, how the results of PISA mathematics might best feed into any review of mathematics education in Ireland; equity in educational opportunity as evident in syllabus take-up; the standards being applied at Junior Certificate at ordinary level in particular (perhaps via the marking schemes applied); the appropriateness of modes of assessment at Junior Certificate; and the need to explore possibilities in Senior Cycle (Transition Year) for students to extend their mathematical problem-solving skills in real-life and cross-curricular contexts. Points 6 to 13, on the other hand, would appear to relate more to theoretical and technical considerations. These pertain to the need to better understand the nature of achievement variance as it relates to sample design. They also concern the importance of gaining a better understanding of how social intake measures relate to achievement – why these appear to vary according the measure used and in some cases, for certain subgroups; and the need to supplement empirical findings with qualitative observations to obtain a more complete picture of what is happening 'on the ground'. The theoretical issues raised also concern possible ways of improving PISA's design to control for non-response bias; and in the absence of alternatives to the present methods, transparency and progressiveness on the part of the OECD on this issue, and increased

efforts by participating countries to enhance response rates. Finally, they concern the potential merits of investigating the effects of extending the existing range of item difficulties included on the PISA tests, particularly at the lower end of the distribution. Some (perhaps all) of the more technical issues raised have clear practical implications regarding the appropriate interpretation of the results. For example, if one is unaware of the methods used to adjust for non-response and the potential for bias, one is more likely to make uninformed and potentially incorrect conclusions based on the PISA results.

In conclusion, it is evident that the technical complexity of international comparative assessments of student achievement has increased. Since the interpretation of international assessments has always been prone to distortion and misrepresentation, perhaps now this is even more the case since methods of scaling the test data and conducting explanatory analyses of achievement are highly technical. In particular, the notion of a probabilistic model of achievement, and the results of explanatory analyses which take account of multiple variables simultaneously, can be difficult to communicate in a clear, concise manner relevant to policy development. Despite these difficulties and complexities (or perhaps precisely because of them!) the results of PISA have, for the most part, been accepted at face value and the subtleties of the findings lost amongst country rankings and absolutist (and at times alarmist) claims about the consequences of the results for the economy.

This thesis aimed to demonstrate that a critical reflection on the interpretability and utility of PISA's results in the Irish context can enhance our understanding of how the results can be used for policy development, and how results should not be used; a secondary result of this exercise is the potential for theoretical development and understanding. The issues raised by Postlethwaite (1999) regarding the need for 'information brokers' of these surveys to act as intermediaries between technically-minded educational researchers and psychometricians on the one hand, and policy-minded members of the government and education practitioners (teachers) on the other to promote informed decisions, would appear to be very relevant.

REFERENCES

Adams, R.J., & Gonzales, E.J. (1996). The TIMSS test design. In M.O. Martin & D.L. Hickey (Eds.), *TIMSS 1995 technical report, Volume I: Design and development* (pp. 3.1-3.26). Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Adams, R.J., Rust, K., & Monseur, C. (2002). Data adjudication. In R. Adams & M. Wu (Eds.), *PISA 2000 technical report* (pp. 180-193). Paris: OECD.

Adams, R.J., & Wu, M. (Eds.) (2002). *PISA 2000 technical report.* Paris: OECD.

Adams, R.J., & Wu, M. (2003). *The impact of item selection on between country comparisons in PISA.* Paper presented at the American Educational Research Association (AERA) Annual Meeting, Chicago, April.

Adams, R.J., Wilson, M.R., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1-24.

Aitkin, M., Francis, B., & Hinde, J. (2005). *Statistical modelling in GLIM 4.* Oxford: Oxford University Press.

Applied Research Branch Strategic Policy, Human Resources Development, Statistics Canada (2000). *Youth in Transition Survey: Project overview.* Quebec: Author. Retrieved July 2005 from http://www.statcan.ca/english/freepub/81-588-XIE/81-588-XIE2000001.pdf

Beaton, A.E., Mullis, V.S., Martin, M.O., Gonzalez, E.J., Kelly, D.L., & Smith, T.A. (1996a). *Mathematics achievement in the middle-school years: IEA's Third International Mathematics and Science Study.* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Beaton, A.E., Mullis, I.V., Martin, M.O., Gonzalez, E.J., Smith, T.A., & Kelly, D.L. (1996b). *Science achievement in the middle-school years: IEA's Third International Mathematics and Science Study.* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Beaton, A.E., Postlethwaite, T.N., Ross, K.N., Spearritt, D., & Wolf, R.M. (1999). *The benefits and limitations of international educational achievement studies.* Paris: International Institute for Educational Planning/UNESCO.

Blum, A., Goldstein, F., & Guérin-Pace, F. (2001). International Adult Literacy Survey (IALS): An analysis of international comparisons of adult literacy. *Assessment in Education, 8,* 225-246.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement, 27,* 165-174.

Bonnet, G. (2002). Reflections in a critical eye: On the pitfalls of international assessment. *Assessment in Education, 9,* 387-400.

Bourdieu, P. (1973). Cultural reproduction and social reproduction. In R. Brown (Ed.), *Papers on the sociology of education* (pp. 71-112). London: Tavistock.

Bourdieu, P. (1984). *Distinction: A social critique of the judgement of taste.* Cambridge, MA: Harvard University Press.

Bottani, N., & Tuijnman, A. (1994). The design of indicator systems. In A.C. Tuijnman & T.N. Postlethwaite (Eds.), *Monitoring the standards of education* (pp. 47-78). Oxford: Pergamon.

Brookover, W.B., Schweitzer, J.H., Schneider, J.M., Beady, C.H., Flood, P.K., & Wisenbaker, J.M. (1978). Elementary school social climate and school achievement. *American Educational Research Journal, 15*, 301-318.

Bryman, A., & Cramer, D. (2004). *Quantitative data analysis with SPSS 12 and 13: A guide for social scientists.* London: Routledge.

Burton, B. (1993). *Some observations on the effect of centering on the results obtained from hierarchical linear modelling.* Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA, April.

Bushnik, T., Barr-Telford, L., & Bussiére, P. (2004). *In and out of high school: First results from the second cycle of the Youth in Transition Survey.* Ottawa: Human Resources Development, Statistics Canada.

Campbell, J.R., Kelly, D.L., Mullis, I.V.S., Martin, M.O., & Sainsbury, M. (2001), *Framework and specifications for PIRLS assessment 2001* (2nd ed.). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Carroll, J. (1963). A model for school learning. *Teachers' College Record, 64*, 723-733.

Dempsey, N. (2004). *Raising the quality of learning for all, 18-19 March 2004, Dublin, Ireland – Chair's Summary.* Retrieved August 2005 from http://www.cmec.ca/international/oecd/OECD.EdMin2004.Chair.en.pdf.

Cizek, G.J. (Ed.) (2001). *Setting performance standards: Concepts, methods, and perspectives.* London: Lawrence Erlbaum Associates.

Close, S., & Oldham, E.E. (2005). Junior Cycle mathematics examinations and the PISA mathematics framework. In Close, S., Dooley, T., & Corcoran, D. (Eds.), *Proceedings of the First National Conference of Research in Mathematics Education* (pp. 174-192). Dublin: St. Patrick's College.

Close, S., Oldham, E.E., Shiel, G., & Cosgrove, J. (2005). *The OECD Programme for International Student Assessment: Interpreting the performance of Irish 15-year olds in mathematics.* Paper presented at European Conference on Educational Research (ECER), Dublin, September.

Cochran, W.G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

Collaer, M.L., & Nelson, J.D. (2002). Large visuospatial sex difference in line judgement: Possible role of attentional factors. *Brain and Cognition, 49,* 1-12.

Coleman, J.S. (1988). Social capital in the creation of human capital. *American Journal of Sociology, 94,* 95-140.

Coleman, J.S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Wienfield, F., & York, R. (1966). *Equality of educational opportunity.* Washington, DC: US Government Printing Office.

Cosgrove, J., Kellaghan, T., Forde, P., & Morgan, M. (2000). *The 1998 national assessment of English reading (with comparative data from the 1993 assessment).* Dublin: Educational Research Centre.

Cosgrove, J., Shiel, G., Sofroniou, N., Zastrutzki, S., & Shortt, F. (2005). *Education for life: The achievements of 15-year-olds in Ireland in the second cycle of PISA.* Dublin: Educational Research Centre.

Cosgrove, J., Shiel, G., Oldham, E.E., & Sofroniou, N. (in preparation). *Analyses of responses on the PISA 2003 mathematics teacher questionnaire.* Dublin: Educational Research Centre.

Coulombe, S., Tremblay, J.S., & Marchand, S. (2004). *International Adult Literacy Survey: Literacy scores, human capital and growth across fourteen OECD countries.* Ottawa: Statistics Canada.

deLange, J. (1994). Curriculum change: an American-Dutch perspective. In D. F. Robitaille, D. H. Wheeler, and C. Kieran (Eds.), *Selected lectures from the 7th International Congress on Mathematical Education* (pp. 229-248). Sante-Foye, Canada: Les Presses de L'Université Laval.

deLange, J. (1998). Real problems with real world mathematics. In C. Alcina, J. Alvarez, M. Niss, A. Pérez, L. Rico, & A. Sfard (Eds.), *Proceedings of the 8th International Congress on Mathematical Education* (pp. 83-110). Seville: S.A.E.M THALES.

Deaton, A. (2002). Policy implications of the gradient of health and wealth. *Health Affairs, 21 (2),* 13-30.

Department of Education and Science (2001, December 4). New OECD report gives major boost to government's literacy measures. *Department of Education and Science Press Release* (retrieved September 2005 from http://www.education.ie/).

Department of Education and Science (2004, December 7). OECD publishes PISA 2003 summary results for Ireland, Irish students rank amongst top in OECD at reading and maintain position in Mathematics and Science. *Department of Education and Science Press Release* (retrieved September 2005 from http://www.education.ie/).

Department of Education and Science (2005, August 25). Minister Hanafin welcomes publication of 'Key Education Statistics'. *Department of Education and Science Press Release* (retrieved September 2005 from http://www.education.ie/).

Department of Education and Science (2005). *Key education statistics – 1993/94-2003/04.* Dublin: Stationery Office. Retrieved August 20005 from http://www.education.ie/

Department for Education and Skills (DfES) (2004). *PISA: Proposals for two international collaborative studies.* Unpublished submission by the DfES to the OECD Directorate for Education, May 14.

Department for Education and Skills (DfES) (2005). *DfES seminar on international studies of educational attainment, draft note.* Unpublished seminar proceedings, March 21.

Department of Education (n.d.). *The Junior Certificate English syllabus.* Dublin: Stationery Office.

Department of Education and Science/National Council for Curriculum and Assessment. (2000). *Mathematics syllabus: Higher, Ordinary, and Foundation Level.* Dublin: Stationery Office.

Department of Education and Science/National Council for Curriculum and Assessment. (2002). *Junior Certificate mathematics: Guidelines for teachers.* Dublin: Stationery Office.

Dunn, O.J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56*, 52-64.

Educational Research Centre (2005, September). *Proceedings of the second national PISA symposium (Regency Hotel, Drumcondra, Dublin).* Dublin: Author.

Eivers, E., Shiel, G., & Shortt, F. (2004). *Reading literacy in disadvantaged primary schools.* Dublin: Educational Research Centre.

Eivers, E., Shiel, G., Perkins, R., & Cosgrove, J. (in preparation). *The 2004 National Assessment of English Reading.* Dublin: Educational Research Centre.

Elley, W.B. (1992). *How in the world do students read? IEA Study of Reading Literacy.* The Hague: International Association for the Evaluation of Educational Achievement.

Floden, R.E. (2002). The measurement of opportunity to learn. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 231-266). Washington, DC: National Academy Press.

Flynn, S. (2004, December 7). Low Irish maths standards criticised. *Irish Times* (retrieved September 2005 from http://www.ireland.com).

Foshay, A.W., Thorndike, R.L., Hotyat, F., Pidgeon, D.A., & Walker, D.A. (1962). *Educational achievements of thirteen-year olds in twelve countries.* Hamburg: UNESCO.

Foy, P. (1998). Implementation of the TIMSS sample design. In M.O. Martin & D.L. Kelly (Eds.), *TIMSS technical report volume II: Implementation and analysis (middle school years).* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Foy, P., Rust, K., & Schleicher, A. (1996). Sample design. In M.O. Martin & D.L. Kelly (Eds.), *TIMSS 1995 technical report, Volume I: Design and development.* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Freudenthal, H. (1973). *Mathematics as an educational task.* Dordrecht: Kluwer Academy.

Freudenthal, H. (1981). Major problems in mathematics education. *Educational Studies in Mathematics, 12,* 133-150.

Gamoran, A.(1992). The variable effects of high school tracking. *Sociology of Education, 57,* 812-828.

Ganzeboom, H.B., & Treiman, D.J. (1996). Internationally comparable measures of occupational status for the 1988 international standard classification of occupations. *Social Science Research, 25,* 201-239.

Gibbons, T., & Sanderson, G. (2002). Contemporary themes in the research enterprise. *International Education Journal, 3(4),* 1-22. Retrieved August 2005 from http://www.flinders.edu.au/education/eij

Goldstein, H. (1987). *Multilevel models in educational and social research.* London: Griffin.

Goldstein, H. (1995). *Interpreting international comparisons of student achievement.* Paris: UNESCO.

Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement, 8,* 369-395.

Goldstein, H. (2000). School effectiveness research and educational policy. *Oxford Review of Education, 26,* 353-363.

Goldstein, H. (2004). International comparisons of student achievement: Some issues arising from the PISA study. *Assessment in Education 11*, 319-330.

Goldstein, H., Hiqui, P., Rath, T., & Hill, N. (1999). The use of value added information in judging school performance. *Report based on a study funded by OFSTED.* Retrieved July 2004 from http://www.ioe.ac.uk/hgpersonal/ value-added-school-performance.html.

Goldstein, H., & Spiegelhalter, D.J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, 159,* 385-443.

Greaney, V., & Kellaghan, T. (1996). *Monitoring the learning outcomes of education systems.* Washington, DC: The World Bank.

Haertel, E.H. (1997). Exploring and explaining US TIMSS performance. *Paper prepared for Learning from TIMSS: An NRC Symposium on the Third International Mathematics and Science Study*, Washington DC, February. Retrieved July 2003 from http://www.enc.org/topics/assessment/ timss/additional/

Hambleton, R.K., Gonzalez, E., Plake, B.S., & Ponocny, I. (2005). *Technical review of PISA: Executive summary, sections 1, 2 and 5* (incomplete and unpublished draft, September 14). Paris: Network A.

Harris, J., Forde, P., Archer, P., Nic Fhearaile, S., & O'Gorman, M. (in preparation). *Irish in primary schools: National trends in achievement (1985-2002).* Unpublished manuscript.

Healy, A. (2003, September 17). Republic fifth in literacy rankings. *Irish Times* (retrieved September 2005 from http://www.ireland.com).

Henderson, V., Mieszkowski, P., & Sauvageau, Y. Peer group effects and educational production functions. *Journal of Public Economics, 10,* 97-106.

Husén, T.N. (Ed.) (1967a). *International study of achievement in mathematics: A comparison of 12 countries (Vol. I).* London: Wiley.

Husén, T.N. (Ed.) (1967b). *International study of achievement in mathematics: A comparison of 12 countries (Vol. II).* London: Wiley.

Husén, T.N. (1973). Foreword. In L.C. Comber & J.P. Keeves (Eds.), *Science achievement in nineteen countries* (pp. 13-24). New York: Wiley.

Husén, T.N. (1997). Research paradigms in education. In J. Keeves (Ed.), *Educational research, methodology, and measurement: An international handbook* (2nd ed.; pp. 17-21). Oxford: Pergamon.

Husén, T.N., & Postlethwaite, T.N. (1996). A brief history of the International Association for the Evaluation of Educational Achievement (IEA). *Assessment in Education, 3,* 129-141.

Husén, T.N., & Tuijnman, A.C. (1994). Monitoring standards in education: Why and how it came about. In A.C. Tuijnman & T.N. Postlethwaite (Eds.), *Monitoring the standards of education* (pp. 1-22). Oxford: Pergamon.

Hutcheson, G., & Sofroniou, N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models.* London: Sage.

International Reading Association (2003). *Policy and practice implications of the Program for International Student Assessment (PISA) 2000.* Newark, DE: Author.

Ireland (1981). *Department of Education annual statistical report 1979-1980.* Dublin: Stationery Office.

Ireland (2001). *Department of Education annual statistical report 1999-2000.* Dublin: Stationery Office.

Irish Times Editorial (2001, December 5). Irish students score second highest marks in Europe for reading ability. *Irish Times* (retrieved September 2005 from http://www.ireland.com).

Irish Times Editorial (2002, January 15). Now is the time to place a proper value on our teachers. *Irish Times* (retrieved September 2005 from http://www.ireland.com).

Irish Times Editorial (2002, October 30). Junior Certificate. *Irish Times* (retrieved September 2005 from http://www.ireland.com).

Irish Times Editorial (2004, December 7). Standards in maths. *Irish Times* (retrieved September 2005 from http://www.ireland.com).

Kaiser, G., Leung, F.K.S., Romberg, T., & Yashenko, I. (2002). International comparisons in mathematics education: An overview. *Proceedings of the International Conference in Mathematics, Beijing, 2002, Vol. I,* pp. 631-646. Retrieved July 2005 from http://arxiv.org/PS_cache/math/pdf/0212/0212416.pdf

Kellaghan, T. (1995). *National monitoring of education in Ireland.* Dublin: Educational Research Centre.

Kellaghan, T. (1996). IEA studies and educational policy. *Assessment in Education, 3,* 143-160.

Kellaghan, T. (2001). Reading literacy standards in Ireland. *Oideas, 49,* 7-20.

Kellaghan, T., & Dwan, B. (1995). *The 1994 Junior Certificate Examination: A review of results.* Dublin: National Council for Curriculum and Assessment.

Kellaghan, T., & Greaney, V. (2001). The globalisation of assessment in the 20th century. *Assessment in Education, 8(1),* 87-102.

Kellaghan, T., & Madaus, G.F. (2000). Outcome evaluation. In D.L. Stufflebeam, G.F. Madaus, & T. Kellaghan (Eds.), *Evaluation models* (pp. 97-112). Boston: Kluwer Academic.

Kellaghan, T., Madaus, G.F., & Rakow. E.A. (1979). Within-school variation in achievement: School effects or error? *Studies in Educational Evaluation , 5,* 101-107.

Kirsch, I., de Jong, J., Lafontaine, D., McQueen, J., Mendelovits, J., & Monseur, C. (2002). *Reading for change: Performance and engagement across countries. Results from PISA 2000.* Paris: OECD.

Kish, L. (1965). *Survey sampling.* New York: Wiley.

Klein, S.P., & Hamilton, L. (1999). *Large-scale testing: Current practices and new directions.* Santa Monica, CA: RAND.

Krawchuk, S., & Rust, K. (2002). Sample design. In R. Adams & M. Wu (Eds.), *PISA 2000 technical report* (pp. 39-56). Paris: OECD.

Lapointe, A.E., Mead, N.A., & Askew, J.M. (1992). *Learning mathematics: The International Assessment of Educational Progress.* Princeton, NJ: ETS.

Lapointe, A.E., Mead, N.A., & Phillips, G.W. (1989). *A world of differences: An international assessment of mathematics and science*. Princeton, NJ: ETS.

Lyons, M., Lynch, K., Close, S., Sheerin, E., & Boland, P. (2003). *Inside classrooms: The teaching and learning of mathematics in social context.* Dublin: Institute of Public Administration.

Madaus, G.F., Kellaghan, T., & Rakow, E.A. (1976). School and class differences in performance on the Leaving Certificate Examination. *Irish Journal of Education, 10,* 41-50.

Madaus, G.F., Airasian, P.W., & Kellaghan, T. (1980). *School effectiveness: A reassessment of the evidence.* New York: McGraw-Hill.

Madaus, G.F., Kellaghan, T., Rakow, E.A., & King, D.J. (1979). The sensitivity of measures of school effectiveness. *Harvard Educational Review, 49,* 207-229.

Martin, M.O., Gregory, K.D., & Stemler, S.E. (2000). *TIMSS 1999 technical report.* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Martin, M.O., & Hickey, B.L. (1992). *1991 Leaving Certificate Examination: A review of results*. Dublin: National Council for Curriculum and Assessment.

Martin, M.O., & Hickey, B.L. (1993). *1992 Junior Certificate Examination: A review of results*. Dublin: National Council for Curriculum and Assessment.

Martin, M.O., & Mullis, I.V.S. (1996). *Quality assurance in data collection: Third International Mathematics and Science Study.* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Martin, M.O., & Mullis, I.V.S. (2000). TIMSS 1999: An overview. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 technical report* (pp. 3-25). Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., & Chrostowski, S.J. (2004). *TIMSS 2003 international science report: Findings From IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., Gregory, K.D., Smith, T.A., Chrostowski, S.J., Garden, R.A., & O'Connor, K.M. (2000a). *TIMSS 1999 international science report: Findings from IEA's repeat of the Third International Mathematics and Science Study at the eighth grade.* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Martin, M.O., Mullis, I.V., Gregory, K.D., Hoyle, C., & Shen, C. (2000b). *IEA's Third International Mathematics and Science Study: Effective schools in science and mathematics.* Boston: TIMSS International Study Center, Boston College.

Mayeske, G.W., Wisler, C.E., Beaton, A.E., Weinfeld, F.D., Cohen, W.M., Okada, D., Proshek, J.M., & Tabler, K.A. (1969). *A study of our nation's schools.* Washington, DC: Department of Health, Education and Welfare, Office of Education.

Millar, D., & Kelly, D. (1999). *From Junior to Leaving Certificate: A longitudinal study of 1994 Junior Certificate candidates who took the Leaving Certificate Examination in 1997.* Dublin: NCCA.

Monseur, C., Rust, K., & Krawchuk, S. (2003). Sampling outcomes. In R.J. Adams & M. Wu (Eds.), *PISA 2000 technical report* (pp. 133-148). Paris: OECD.

Monseur, C., & Wu, M. (2002). Imputation for student non-response in educational achievement surveys. Paper presented at the International Conference on Improving Surveys (25-28 August, Copenhagen, Denmark, 2002). Retrieved February 2004 from http://www.icis.dk/ICIS_papers/E2_5_2.pdf

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., & Chrostowski, S.J. (2004). *TIMSS 2003 international mathematics report: Findings From IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., Gregory, K.D., Garden, R.A., O'Connor, K.M., Chrostowski, S.J., & Smith, T.A. (2000). *TIMSS 1999 international mathematics report: Findings from IEA's repeat of the Third International Mathematics and Science Study at the eighth grade.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., & Kennedy, A.M. (2003a). PIRLS 2001 international report: IEA's study of reading literacy achievement in primary schools. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., & Flaherty, C.L. (2002). *PIRLS 2001 encyclopaedia: A reference guide to reading education in the countries participating in IEA's Progress in International Reading Literacy Study (PIRLS).* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I.V.S., Martin, M.O., Smith, T.A., Garden, R.A., Gregory, K.D., Gonzalez, E.J., Chrostowski, S.J., & O'Connor, K.M. (2003b). *TIMSS: Assessment framework and specifications 2003* (2nd ed.). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

McCloughlin, T. (2005). *Timeo Danaos:* An integrated mathematics and science programme for Transition Year. In S. Close, T. Dooley, & D. Corcoran (Eds.), *Proceedings of the First National Conference of Research in Mathematics Education* (pp. 209-219). Dublin: St. Patrick's College.

Nash, R. (2003). Is the school composition effect real? A discussion with evidence from the UK PISA data. *School Effectiveness and School Improvement, 14,* 441-457.

National Academy of Education (NAE). (1993). *Setting performance standards for student achievement: A report of the National Academy of Education Panel on the Evaluation of the NAEP trial state assessment: An evaluation of 1992 achievement levels.* Stanford, CA: Author.

National Council for Curriculum and Assessment (NCCA). (1989). *A guide to the Junior Certificate.* Dublin: Author.

National Council for Curriculum and Assessment (NCCA). (2004, April). *Update on the junior cycle review.* Dublin: Author.

National Council for Curriculum and Assessment (NCCA). (2005, October). *Review of mathematics in post-primary education: A discussion paper.* Dublin: Author.

National Council on Education Standards and Testing (NCEST). (1992). *Raising standards for American education.* Washington, DC: US Government Printing Office.

National Commission on Excellence in Education (NCEE) (1983). *A nation at risk. The imperative for educational reform.* Washington, DC: Author.

Nohara, D. (2001). *A comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA).* Washington, DC: National Center for Education Statistics.

OECD (Organisation for Economic Co-operation and Development). (1992a). *Education at a glance: OECD indicators.* Paris: Author.

OECD (Organisation for Economic Co-operation and Development). (1992b). *The OECD international education indicators: A framework for analysis.* Paris: Author.

OECD (Organisation for Economic Co-operation and Development). (1998). *Human capital investment: An international comparison.* Paris: Author.

OECD (Organisation for Economic Co-operation and Development). (1999a). *Classifying educational programmes: Manual for ISCED-97 implementation in OCED countries.* Paris: Author.

OECD (Organisation for Economic Co-operation and Development). (1999b). *Education at a glance: OECD indicators.* Paris: Author.

OECD (Organisation for Economic Co-operation and Development). (1999c). *Measuring student knowledge and skills: A new framework for assessment.* Paris: Author.

OECD (Organisation for Economic Co-operation and Development). (2000a). *Education at a glance: OECD indicators.* Paris: Author.

OECD (Organisation for Economic Co-operation and Development). (2000b). *Measuring student knowledge and skills: The PISA 2000 assessment of reading, mathematical and scientific literacy.* Paris: Author.

OECD (Organisation for Economic Co-operation and Development). (2001a). *Education at a glance: OECD indicators.* Paris: Author.

OECD (Organisation for Economic Co-operation and Development).. (2001b). *Knowledge and skills for life: First results of PISA 2000.* Paris: Author.

OECD (Organisation for Economic Co-operation and Development). (2002a). *Education at a glance: OECD indicators.* Paris: Author.

OECD (Organisation for Economic Co-operation and Development). (2002b). *Manual for the PISA 2000 database.* Paris: Author.

OECD (Organisation for Economic Co-operation and Development). (2002c). *Sample tasks from the PISA 2000 assessment: Reading, mathematical and scientific literacy.* Paris: Author.

OECD (Organisation for Economic Co-operation and Development). (2003b). *The PISA 2003 assessment framework: Mathematics, reading, science and problem-solving knowledge and skills.* Paris: Author.

OECD (Organisation for Economic Co-operation and Development). (2003a). *Learners for life – student approaches to learning: Results from PISA 2000.* Paris: Author.

OECD (Organisation for Economic Co-operation and Development). (2004a). *Completing the foundation for lifelong learning: An OECD survey of upper secondary schools.* Paris: Author.

OECD (Organisation for Economic Co-operation and Development). (2004b). *Education at a glance: OECD indicators.* Paris: Author.

OECD (Organisation for Economic Co-operation and Development). (2004c). *Learning for tomorrow's world: First results from PISA 2003.* Paris: Author.

OECD (Organisation for Economic Co-operation and Development).. (2004d). *Problem-solving for tomorrow's world: First measures of cross-curricular skills from PISA 2003.* Paris: Author.

OECD (Organisation for Economic Co-operation and Development). (2005a). *PISA 2003 data analysis manual (SPSS users).* Paris: Author.

OECD (Organisation for Economic Co-operation and Development). (2005b). *PISA 2003 technical report.* Paris: Author.

OECD (Organisation for Economic Co-operation and Development). (2005c). *School factors related to quality and equity: Results from PISA 2000.* Paris: Author.

OECD (Organisation for Economic Co-operation and Development)/HDRC (Human Resources Development, Canada) (1997). *Literacy skills for the knowledge society: Results from the First International Adult Literacy Survey.* Paris: OECD.

OECD (Organisation for Economic Co-operation and Development) Secretariat (2005, February 17). *Nineteenth meeting of PISA Governing Board: Draft agenda.* Unpublished meeting document of the PISA Governing Board. Paris: Author.

OECD (Organisation for Economic Co-operation and Development) Secretariat (2005, September 26). *Longer-term strategy of the development of PISA.* Unpublished meeting document of the PISA Governing Board. Paris: Author.

Oldham, E. E. (1980a). Case studies in algebra education: Ireland. In *Proceedings of the joint IDM/IEA conference: Comparative studies of mathematics curricula: Change and stability 1960-1980* (pp. 326-346). Bielefeld: Institut für Didaktik der Mathematik der Universität Bielefeld.

Oldham, E. E. (1980b). Case studies in geometry education: Ireland. In *Proceedings of the joint IDM/IEA conference: Comparative studies of mathematics curricula: Change and stability 1960-1980* (pp. 395-425). Bielefeld: Institut für Didaktik der Mathematik der Universität Bielefeld.

Oldham, E. E. (1989). Is there an international mathematics curriculum? In B. Greer & G. Mulhern (Eds.), *New directions in mathematics education* (pp. 185-224). London: Routledge.

Oldham, E. E. (2001). The culture of mathematics education in the Republic of Ireland: Keeping the faith? *Irish Educational Studies*, *20*, 266-277.

Oldham, E. E. (2002). The performance of Irish students in mathematical literacy in the Programme for International Student Assessment (PISA). *Irish Journal of Education, 33,* 31-52.

Oliver, E. (2001, December 5). Irish 15-year-olds show up well in international education survey. *Irish Times* (retrieved September 2005 from http://www.ireland.com).

Oliver, E. (2002, October 30). Pupils impress and report notes low education spend. *Irish Times* (retrieved September 2005 from http://www.ireland.com).

O'Donoghue, C., Thomas, S., Goldstein, H., & Knight, T. (1997). 1996 DfEE study of value-added for 16-18 year olds in England. *DfEE Research Series, March, 1997.* London: DfEE. Retrieved June 2005 from http://www.mlwin.com/hgpersonal/alevdfee.pdf

O'Leary, M. (2001). The effects of age-based and grade-based sampling on the relative standing of countries in international comparative studies of student achievement in science. *British Educational Research Journal, 27,* 187-200.

O'Leary, M., Madaus, G.F., Kellaghan, T., & Beaton, A. (2000). The consistency of findings across international surveys of mathematics and science achievement: A comparison of IAEP2 and TIMSS. *Education Policy Analysis Archives, 8(43).* Retrieved June 2004 from http://epaa.asu.edu/epaav8n43.html

Osborne, J.W. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research & Evaluation, 7(1).* Retrieved June 2005 from http://PAREonline.net/getvn.asp?v=7&n=1

PISA Consortium (2000a). PISA 2000: Marking guide for mathematics. Unpublished procedural document for use by national centres. Melbourne: Australian Council for Educational Research.

PISA Consortium (2000b). PISA 2000: Marking guide for reading. Unpublished procedural document for use by national centres. Melbourne: Australian Council for Educational Research.

PISA Consortium (2003a). PISA 2003: Marking guide for mathematics. Unpublished procedural document for use by national centres. Melbourne: Australian Council for Educational Research.

PISA Consortium (2003b). PISA 2003: Marking guide for reading. Unpublished procedural document for use by national centres. Melbourne: Australian Council for Educational Research.

Plomp, T., Howie, S., & McGaw, B. (2003). International studies of educational achievement. In T. Kellaghan & D.L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 951-978). Dordrecht: Kluwer Academic.

Postlethwaite, T.N. (1995). Calculation and interpretation of between-school and within-school variance in achievement. In OECD, *Measuring what students learn* (pp. 81-91). Paris: OECD.

Postlethwaite, T.N. (1999). *International studies of educational achievement: Methodological issues. CERC studies in comparative education 6.* Hong Kong: Comparative Education Research Centre, University of Hong Kong.

Prais, S. J. (2003). Cautions on OECD's recent education survey (PISA). *Oxford Review of Education, 29,* 139-163.

Purves, A.C. (1975). *Educational policy and international assessment: Implications of the IEA surveys of achievement.* Berkely, CA: McCutchan Publishing Foundation.

Raudenbush, S.W., & Bryk, A.S (1992). *Hierarchical linear models: Applications and data analysis.* Newbury Park, CA: Sage.

Raudenbush, S.W., Bryk, A.S., Cheong, Y.F., & Congdon, R.T. (2004). *HLM 6: Hierarchical linear and non-linear modelling.* Lincolnwood, IL: Scientific Software International, Inc.

Raudenbush, S.W., & Willms, J.D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20,* 307-335.

Robitaille, D.F., & Garden, R.A. (1989). *The IEA Study of Mathematics II: Contexts and outcomes of school mathematics.* Oxford: Pergamon.

Robitaille, D.F., Schmidt, W.H., Raizen, S., McKnight, C., Britton, E., & Nicol, C. (1993). *Curriculum frameworks for mathematics and science: the Third International Mathematics and Science Study.* Vancouver: Pacific Educational Press.

Robitaille, D.F., & Travers, K.J. (1992). International studies of achievement in mathematics. In D.A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 687-709). New York: Macmillan.

Rothman, R. (2002, March). *The impact of international studies in education: The view of education journalists.* Unpublished document. Paper prepared for the Board on International Comparative Studies in Education, National Research Council, available from the author (robert_rothman@brown.edu).

Rust, K., & Krawchuk, S. (2002), Survey weighting and the calculation of sampling variance. In R.J. Adams & M. Wu (eds.), *PISA 2000 technical report* (pp. 89-98). Paris: OECD.

Schulz, W. (2002). Constructing and validating the questionnaire indices. In R.J. Adams & M. Wu (Eds.), *PISA 2000 technical report* (pp. 217-252). Paris: OECD.

Shiel, G., Cosgrove, J., Sofroniou, N., & Kelly, A. (2001). *Ready for life: The literacy achievements of Irish 15-year olds.* Dublin: Educational Research Centre.

Shiel, G., & Kelly, D. (2001). *The 1999 national assessment of mathematics achievement.* Dublin: Educational Research Centre.

Shiel, G., & Surgenor, P. (in preparation). *Report on the 2004 National Assessment of Mathematics Achievement.* Dublin: Educational Research Centre.

Smithers, A. (2004). *England's education: What can be learned by comparing countries? Report to the Sutton Trust, May 2004.* Liverpool: Centre for Education and Employment Research. Retrieved June 8, 2005 from http://www.suttontrust.com/reports/pisa_publication.doc

Smyth, E. (1999). *Do schools differ? Academic and personal development among pupils in the second-level sector.* Dublin: ESRI/Oak Tree Press.

Smyth, E., Dunne, A., McCoy, S., & Darmody, M. (in preparation). *Pathways through the Junior Cycle: The experiences of second year students.* Dublin: ESRI.

Smyth, E., McCoy, E., & Darmody, M. (2004). *Moving up: The experiences of first year students in post-primary education.* Dublin: Liffey Press in association with the ESRI.

Snijders, A.B., & Bosker, R.J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modelling.* London: Sage.

Sofroniou, N., Archer, P., & Weir, S. (in preparation, authorship subject to revision). *An analysis of the association between socioeconomic context, gender and achievement.* Dublin: Educational Research Centre.

Sofroniou, N., Cosgrove, J., & Shiel, G (2002). Using PISA variables to explain performance on the Junior Certificate Examinations in mathematics and science. *Irish Journal of Education, 33,* 99-124.

Sofroniou, N., Shiel, G., & Cosgrove, J. (2000). Explaining performance on the Junior Certificate Examination: A multilevel approach. *Irish Journal of Education, 31,* 25-49.

Sofroniou, N., Shiel, G., & Cosgrove, J. (2002). PISA reading literacy in Ireland: An expanded model exploring attributes of self-regulated learning. *Irish Journal of Education, 33,* 97-98.

Spaulding, S. (1989). Comparing educational phenomena. In A.C. Purves (Ed.), *International comparisons and educational reform* (pp. 1-16). Washington, DC: Association for Supervision and Curriculum Development.

Thrupp, M. (1999). *Schools making a difference: Let's be realistic! Social mix, school effectiveness, and the social limits of school reform.* Balmoor, Bucks.: Open University Press.

Treffers, A. (1987). *Three dimensions: A model of goal and theory description in mathematics instruction – the Wiskobas project.* Dordrecht: Kluwer Academic.

Turmo, A. (2001). Science achievement and socio-economic status: A discussion based on PISA 2000. *Paper presented at the ESERA conference in Thessalonki, Greece, August 2001.* Retrieved June 2005 from http://www.pisa.no/ Dokumenter/ESERApaperAre.pdf

Turner, R. (2002). Proficiency scales construction. In In R.J. Adams & M. Wu (Eds.), *PISA 2000 technical report* (pp. 195-216). Paris: OECD.

UNESCO (United Nations Educational, Scientific and Cultural Organisation). (2003). *UNESCO: What it is, what it does.* Paris: Author.

Volodin, N., Macaskill, G., Monseur, C., Adams, R.J., & Wu, M. (2003). *KeyQuest version 3 manual.* Unpublished document, PISA International Consortium.

Voyer, D., Voyer, S., & Bryden, M.P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin, 117,* 250-270.

Walshe, J. (2001, December 5). No love for school but top marks for reading. *Irish Independent* (retrieved September 2005 from http://www.unison. independent.ie).

Walshe, J., & Donnelly, K. (2003, September 17). Top class pupils could do better on funding. *Irish Independent* (retrieved September 2005 from http://www.unison. independent.ie).

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427-445.

Watson, N.V., & Kimura, D. (1991). Nontrivial sex differences in throwing and intercepting: Relation to psychometrically-defined spatial functions. *Personality and Individual Differences, 5,* 375-385.

Welch, F. (1999). The triumph of technocracy or the collapse of certainty? Modernity, postmodernity and postcolonialism in comparative education. In F. Arnove & C. Torres (Eds.), *Comparative education: The dialectic of the global and the local* (pp. 25-49). Lanham, MD: Rowman & Littlefield.

Westat. (2000). *WesVar complex samples 4.0.* Rockville, MD: Author.

Westbury, I. (1989). The problems of comparing curriculums across educational systems. In A.C. Purves (Ed.), *International comparisons and educational reform* (pp. 17-34). Washington, DC: Association for Supervision and Curriculum Development.

Wiliam, D. (2004). *The validity and impact of educational assessments.* Paper presented at the 10th International Conference of Mathematics Education, Denmark, July.

Wilson, L., & Zhang, L.R. (1998). *A cognitive analysis of gender differences on constructed-response and multiple-choice assessments of mathematics.* Paper presented at the American Educational Research Association (AERA) Annual Meeting, San Diego, CA, April.

Willms, J.D. (2002). *Ten hypotheses about socioeconomic gradients and community differences in children's developmental outcomes.* Montreal, Quebec: Statistics Canada.

Wu, M. (2002). Test design and test development. In R.J. Adams & M. Wu (Eds.), *PISA 2000 technical report* (pp. 21-31). Paris: OECD.

Wu, M., & Adams, R.J. (2003). *Modelling mathematics problem-solving items using a multidimensional IRT model.* Paper presented at the American Educational Research Association (AERA) Annual Meeting, Chicago, April.

Wu, M., Adams, R.J., & Wilson, M.R. (1997). *ConQuest: Multi-aspect test software.* Camberwell: Australian Council for Educational Research.

Zabulionis, A. (2001). Mathematics and science achievement of various nations. *Education Policy Analysis Archives*, *9*(33). Retrieved March 2004 from http://epaa.asu.edu/epaa/v9n33/

Zhang, L.R., Wilson, L., & Manon, J. (1999). *An analysis of gender differences on performance assessment in mathematics – a follow-up study.* Paper presented at the American Educational Research Association (AERA) Annual Meeting, Montreal, April.

APPENDIX 4: ADDITIONAL TABLES

Table A4.1. Means, Standard Errors, and 95% Confidence Intervals For the Combined Reading Score, and for the Three Reading Process Subscales, For Each Point on the EJCPS: Scale 1 - Sub-Cohort Taking the Junior Certificate English Examination in 2000

EJCPS	N	%	Overall	SE	CI95L	CI95U	Retrieve	SE	CI95L	CI95U	Interpret	SE	CI95L	CI95U	Reflect	SE	CI95L	CI95U
1*	0	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
2	3	0.1	**299.4**	25.33	249.0	349.8	**309.0**	29.23	250.8	367.1	**328.4**	27.36	274.0	382.9	**340.0**	34.47	271.4	408.6
3	5	0.2	**316.1**	25.00	266.4	365.9	**296.3**	47.20	202.4	390.2	**296.7**	36.27	224.5	368.9	**346.5**	27.52	291.7	401.2
4	22	0.9	**331.2**	13.49	304.4	358.1	**329.2**	16.74	295.9	362.5	**333.2**	14.16	305.0	361.4	**343.5**	15.54	312.5	374.4
5	18	0.8	**347.8**	21.62	304.8	390.8	**339.3**	21.71	296.1	382.5	**337.9**	21.25	295.6	380.2	**363.7**	20.71	322.5	404.9
6	155	6.6	**401.0**	6.87	387.3	414.7	**395.3**	8.39	378.6	412.0	**399.6**	8.008	383.6	415.5	**414.6**	8.30	398.1	431.1
7	353	15.0	**439.7**	5.48	428.8	450.6	**432.3**	5.93	420.5	444.1	**437.3**	5.805	425.7	448.8	**452.5**	5.80	441.0	464.0
8	225	9.6	**470.8**	4.62	461.6	480.0	**468.3**	4.88	458.6	478.1	**469.0**	4.784	459.5	478.5	**483.1**	4.96	473.2	493.0
9	426	18.0	**516.9**	4.14	508.6	525.1	**513.4**	4.79	503.9	523.0	**515.1**	4.261	506.6	523.6	**522.6**	4.49	513.7	531.5
10	677	28.7	**552.6**	2.91	546.8	558.4	**551.4**	3.52	544.4	558.4	**553.8**	3.049	547.7	559.9	**560.9**	2.74	555.5	566.4
11	372	15.8	**595.0**	3.77	587.5	602.5	**591.5**	4.50	582.5	600.5	**595.4**	4.63	586.2	604.6	**598.5**	4.31	589.9	607.1
12	103	4.4	**635.3**	6.27	622.9	647.8	**632.5**	6.16	620.2	644.8	**639.7**	5.98	627.8	651.6	**637.9**	6.39	625.1	650.6
All available	2360	100.0	**517.3**	3.56	510.3	524.4	**513.8**	3.71	506.5	521.2	**516.9**	3.67	509.6	524.2	**525.7**	3.44	518.9	532.6

*No student was awarded Grade F at Foundation Level

Table A4.2. Means, Standard Errors, and 95% Confidence Intervals For the Combined Reading Score, and for the Three Reading Process Subscales, For Each Point on the EJCPS: Scale 2 - Sub-Cohort Taking the Junior Certificate English Examination in 2000

EJCPS	N	%	Overall	SE	CI95L	CI95U	Retrieve	SE	CI95L	CI95U	Interpret	SE	CI95L	CI95U	Reflect	SE	CI95L	CI95U
1	30	1.3	**325.7**	11.13	303.6	347.9	**321.7**	13.89	294.0	349.3	**326.4**	12.75	301.1	351.8	**343.7**	12.74	318.3	369.0
2	18	0.8	**347.8**	21.62	304.8	390.8	**339.3**	21.71	296.1	382.5	**337.9**	21.25	295.6	380.2	**363.7**	20.71	322.5	404.9
3	155	6.6	**401.0**	6.87	387.3	414.7	**395.3**	8.39	378.6	412.0	**399.6**	8.008	383.6	415.5	**414.6**	8.30	398.1	431.1
4	353	15.0	**439.7**	5.48	428.8	450.6	**432.3**	5.93	420.5	444.1	**437.3**	5.805	425.7	448.8	**452.5**	5.80	441.0	464.0
5	225	9.6	**470.8**	4.62	461.6	480.0	**468.3**	4.88	458.6	478.1	**469.0**	4.784	459.5	478.5	**483.1**	4.96	473.2	493.0
6	426	18.0	**516.9**	4.14	508.6	525.1	**513.4**	4.79	503.9	523.0	**515.1**	4.261	506.6	523.6	**522.6**	4.49	513.7	531.5
7	677	28.7	**552.6**	2.91	546.8	558.4	**551.4**	3.52	544.4	558.4	**553.8**	3.049	547.7	559.9	**560.9**	2.74	555.5	566.4
8	372	15.8	**595.0**	3.77	587.5	602.5	**591.5**	4.50	582.5	600.5	**595.4**	4.63	586.2	604.6	**598.5**	4.31	589.9	607.1
9	103	4.4	**635.3**	6.27	622.9	647.8	**632.5**	6.16	620.2	644.8	**639.7**	5.98	627.8	651.6	**637.9**	6.39	625.1	650.6
All available	2360	100.0	**517.3**	3.56	510.3	524.4	**513.8**	3.71	506.5	521.2	**516.9**	3.67	509.6	524.2	**525.7**	3.44	518.9	532.6

Table A4.3. Means, Standard Errors, and 95% Confidence Intervals For the Combined Reading Score, and for the Three Reading Process Subscales, For Each Point on the EJCPS: Scale 3 - Sub-Cohort Taking the Junior Certificate English Examination in 2000

EJCPS	N	%	Overall	SE	CI95L	CI95U	Retrieve	SE	CI95L	CI95U	Interpret	SE	CI95L	CI95U	Reflect	SE	CI95L	CI95U
1	48	2.0	**334.0**	11.77	310.6	357.5	**328.3**	11.38	305.7	350.9	**330.8**	10.93	309.0	352.5	**351.2**	11.88	327.6	374.9
2	155	6.6	**401.0**	6.87	387.3	414.7	**395.3**	8.39	378.6	412.0	**399.6**	8.008	383.6	415.5	**414.6**	8.30	398.1	431.1
3	353	15.0	**439.7**	5.48	428.8	450.6	**432.3**	5.93	420.5	444.1	**437.3**	5.805	425.7	448.8	**452.5**	5.80	441.0	464.0
4	225	9.6	**470.8**	4.62	461.6	480.0	**468.3**	4.88	458.6	478.1	**469.0**	4.784	459.5	478.5	**483.1**	4.96	473.2	493.0
5	426	18.0	**516.9**	4.14	508.6	525.1	**513.4**	4.79	503.9	523.0	**515.1**	4.261	506.6	523.6	**522.6**	4.49	513.7	531.5
6	677	28.7	**552.6**	2.91	546.8	558.4	**551.4**	3.52	544.4	558.4	**553.8**	3.049	547.7	559.9	**560.9**	2.74	555.5	566.4
7	372	15.8	**595.0**	3.77	587.5	602.5	**591.5**	4.50	582.5	600.5	**595.4**	4.63	586.2	604.6	**598.5**	4.31	589.9	607.1
8	103	4.4	**635.3**	6.27	622.9	647.8	**632.5**	6.16	620.2	644.8	**639.7**	5.98	627.8	651.6	**637.9**	6.39	625.1	650.6
All available	2360	100.0	**517.3**	3.56	510.3	524.4	**513.8**	3.71	506.5	521.2	**516.9**	3.67	509.6	524.2	**525.7**	3.44	518.9	532.6

Table A4.4. Means, Standard Errors, and 95% Confidence Intervals For the Combined Reading Score, and for the Three Reading Process Subscales, For Each Point on the EJCPS: Scale 4 - Sub-Cohort Taking the Junior Certificate English Examination in 2000

EJCPS	N	%	Overall	SE	CI95L	CI95U	Retrieve	SE	CI95L	CI95U	Interpret	SE	CI95L	CI95U	Reflect	SE	CI95L	CI95U
1	30	1.3	**325.7**	11.13	303.6	347.9	**321.7**	13.89	294.0	349.3	**326.4**	12.75	301.1	351.8	**343.7**	12.74	318.3	369.0
2	173	7.3	**395.4**	6.66	382.2	408.7	**389.5**	7.73	374.1	404.8	**393.1**	7.567	378.0	408.2	**409.3**	7.74	393.9	424.7
3	353	15.0	**439.7**	5.48	428.8	450.6	**432.3**	5.93	420.5	444.1	**437.3**	5.805	425.7	448.8	**452.5**	5.80	441.0	464.0
4	225	9.6	**470.8**	4.62	461.6	480.0	**468.3**	4.88	458.6	478.1	**469.0**	4.784	459.5	478.5	**483.1**	4.96	473.2	493.0
5	426	18.0	**516.9**	4.14	508.6	525.1	**513.4**	4.79	503.9	523.0	**515.1**	4.261	506.6	523.6	**522.6**	4.49	513.7	531.5
6	677	28.7	**552.6**	2.91	546.8	558.4	**551.4**	3.52	544.4	558.4	**553.8**	3.049	547.7	559.9	**560.9**	2.74	555.5	566.4
7	372	15.8	**595.0**	3.77	587.5	602.5	**591.5**	4.50	582.5	600.5	**595.4**	4.63	586.2	604.6	**598.5**	4.31	589.9	607.1
8	103	4.4	**635.3**	6.27	622.9	647.8	**632.5**	6.16	620.2	644.8	**639.7**	5.98	627.8	651.6	**637.9**	6.39	625.1	650.6
All available	2360	100.0	**517.3**	3.56	510.3	524.4	**513.8**	3.71	506.5	521.2	**516.9**	3.67	509.6	524.2	**525.7**	3.44	518.9	532.6

Table A4.5. Means, Standard Errors, and 95% Confidence Intervals For the Combined Reading Score, and for the Three Reading Process Subscales, For Each Point on the EJCPS: Scale 5 - Sub-Cohort Taking the Junior Certificate English Examination in 2000

EJCPS	N	%	Overall	SE	CI95L	CI95U	Retrieve	SE	CI95L	CI95U	Interpret	SE	CI95L	CI95U	Reflect	SE	CI95L	CI95U
1	48	2.0	**334.0**	11.77	310.6	357.5	**328.3**	11.38	305.7	350.9	**330.8**	10.93	309.0	352.5	**351.2**	11.88	327.6	374.9
2	155	6.6	**401.0**	6.87	387.3	414.7	**395.3**	8.39	378.6	412.0	**399.6**	8.008	383.6	415.5	**414.6**	8.30	398.1	431.1
3	352	14.9	**439.5**	5.50	428.6	450.5	**432.2**	5.93	420.4	444.0	**437.1**	5.81	425.5	448.6	**452.4**	5.81	440.8	463.9
4	197	8.3	**471.5**	4.73	462.1	480.9	**468.2**	5.21	457.8	478.6	**469.4**	5.087	459.2	479.5	**483.3**	5.00	473.3	493.2
5	33	1.4	**507.9**	10.92	486.2	529.6	**506.2**	13.47	479.4	533.0	**505.2**	12.9	479.6	530.9	**520.5**	10.54	499.5	541.5
6	1	0.0	**509.6**	15.94	477.9	541.3	**445.6**	41.80	362.4	528.8	**513.2**	27.19	459.1	567.3	**505.2**	20.18	465.1	545.4
7	29	1.2	**465.8**	12.68	440.5	491.0	**469.3**	13.75	441.9	496.6	**466.7**	13.51	439.8	493.6	**481.6**	12.55	456.6	506.6
8	392	16.6	**517.6**	4.29	509.1	526.2	**514.1**	4.92	504.3	523.8	**515.9**	4.45	507.1	524.8	**522.8**	4.82	513.2	532.4
9	677	28.7	**552.6**	2.91	546.8	558.4	**551.4**	3.52	544.4	558.4	**553.8**	3.049	547.7	559.9	**560.9**	2.74	555.5	566.4
10	372	15.8	**595.0**	3.77	587.5	602.5	**591.5**	4.50	582.5	600.5	**595.4**	4.63	586.2	604.6	**598.5**	4.31	589.9	607.1
11	103	4.4	**635.3**	6.27	622.9	647.8	**632.5**	6.16	620.2	644.8	**639.7**	5.98	627.8	651.6	**637.9**	6.39	625.1	650.6
All available	2360	100.0	**517.3**	3.56	510.3	524.4	**513.8**	3.71	506.5	521.2	**516.9**	3.67	509.6	524.2	**525.7**	3.44	518.9	532.6

Table A4.6. Means, Standard Errors, and 95% Confidence Intervals For the Combined Reading Score, and for the Three Reading Process Subscales, For Each Point on the EJCPS: Scale 6 - Sub-Cohort Taking the Junior Certificate English Examination in 2000

EJCPS	N	%	Overall	SE	CI95L	CI95U	Retrieve	SE	CI95L	CI95U	Interpret	SE	CI95L	CI95U	Reflect	SE	CI95L	CI95U
1	48	2.0	**334.0**	11.77	310.6	357.5	**328.3**	11.38	305.7	350.9	**330.8**	10.93	309.0	352.5	**351.2**	11.88	327.6	374.9
2	155	6.6	**401.0**	6.87	387.3	414.7	**395.3**	8.39	378.6	412.0	**399.6**	8.008	383.6	415.5	**414.6**	8.30	398.1	431.1
3	352	14.9	**439.5**	5.50	428.6	450.5	**432.2**	5.93	420.4	444.0	**437.1**	5.81	425.5	448.6	**452.4**	5.81	440.8	463.9
4	197	8.3	**471.5**	4.73	462.1	480.9	**468.2**	5.21	457.8	478.6	**469.4**	5.087	459.2	479.5	**483.3**	5.00	473.3	493.2
5	34	1.4	**507.9**	10.77	486.5	529.4	**504.7**	13.52	477.8	531.6	**505.4**	12.86	479.8	531.0	**520.1**	10.43	499.3	540.9
6	29	1.2	**465.8**	12.68	440.5	491.0	**469.3**	13.75	441.9	496.6	**466.7**	13.51	439.8	493.6	**481.6**	12.55	456.6	506.6
7	392	16.6	**517.6**	4.29	509.1	526.2	**514.1**	4.92	504.3	523.8	**515.9**	4.45	507.1	524.8	**522.8**	4.82	513.2	532.4
8	677	28.7	**552.6**	2.91	546.8	558.4	**551.4**	3.52	544.4	558.4	**553.8**	3.049	547.7	559.9	**560.9**	2.74	555.5	566.4
9	372	15.8	**595.0**	3.77	587.5	602.5	**591.5**	4.50	582.5	600.5	**595.4**	4.63	586.2	604.6	**598.5**	4.31	589.9	607.1
10	103	4.4	**635.3**	6.27	622.9	647.8	**632.5**	6.16	620.2	644.8	**639.7**	5.98	627.8	651.6	**637.9**	6.39	625.1	650.6
All available	2360	100.0	**517.3**	3.56	510.3	524.4	**513.8**	3.71	506.5	521.2	**516.9**	3.67	509.6	524.2	**525.7**	3.44	518.9	532.6

Table A4.7. Means, Standard Errors, and 95% Confidence Intervals For the Combined Reading Score, and for the Three Reading Process Subscales, For Each Point on the EJCPS: Scale 7 - Sub-Cohort Taking the Junior Certificate English Examination in 2000

EJCPS	N	%	Overall	SE	CI95L	CI95U	Retrieve	SE	CI95L	CI95U	Interpret	SE	CI95L	CI95U	Reflect	SE	CI95L	CI95U
1	48	2.0	**334.0**	11.77	310.6	357.5	**328.3**	11.38	305.7	350.9	**330.8**	10.93	309.0	352.5	**351.2**	11.88	327.6	374.9
2	155	6.6	**401.0**	6.87	387.3	414.7	**395.3**	8.39	378.6	412.0	**399.6**	8.008	383.6	415.5	**414.6**	8.30	398.1	431.1
3	352	14.9	**439.5**	5.50	428.6	450.5	**432.2**	5.93	420.4	444.0	**437.1**	5.81	425.5	448.6	**452.4**	5.81	440.8	463.9
4	198	8.4	**471.7**	4.71	462.3	481.1	**468.1**	5.16	457.8	478.4	**469.6**	5.049	459.5	479.6	**483.4**	4.98	473.5	493.3
5	62	2.6	**488.4**	9.30	469.9	506.9	**489.1**	10.56	468.1	510.1	**487.4**	10.14	467.2	507.6	**502.5**	9.40	483.8	521.2
6	392	16.6	**517.6**	4.29	509.1	526.2	**514.1**	4.92	504.3	523.8	**515.9**	4.45	507.1	524.8	**522.8**	4.82	513.2	532.4
7	677	28.7	**552.6**	2.91	546.8	558.4	**551.4**	3.52	544.4	558.4	**553.8**	3.049	547.7	559.9	**560.9**	2.74	555.5	566.4
8	372	15.8	**595.0**	3.77	587.5	602.5	**591.5**	4.50	582.5	600.5	**595.4**	4.63	586.2	604.6	**598.5**	4.31	589.9	607.1
9	103	4.4	**635.3**	6.27	622.9	647.8	**632.5**	6.16	620.2	644.8	**639.7**	5.98	627.8	651.6	**637.9**	6.39	625.1	650.6
All available	2360	100.0	**517.3**	3.56	510.3	524.4	**513.8**	3.71	506.5	521.2	**516.9**	3.67	509.6	524.2	**525.7**	3.44	518.9	532.6

Table A4.8. Pearson Correlations Between Eight EJCPS Scales and PISA 2000 Reading: Combined Scale and Process Subscales - Sub-Cohort Taking the Junior Certificate English Examination in 2000

	EJCPS Scale							
PISA Scale	Scale 1	Scale 2	Scale 3	Scale 4	Scale 5	Scale 6	Scale 7	Scale 8
Combined Scale	.736	.736	.734	.735	.711	.721	.730	.736
Retrieve	.709	.709	.707	.708	.687	.696	.704	.709
Interpret	.722	.723	.720	.722	.699	.708	.717	.723
Reflect	.725	.726	.724	.725	.700	.709	.719	.726

Correlations are all significant (p<.001).

Table A4.9. Means, Standard Errors, and 95% Confidence Intervals For the Combined Reading Score, PISA 2003 Sample (All Students and the Sub-Sample Attempting the Junior Certificate Examination in 2003)

	PISA 2003 - All Participating						PISA 2003 - JCE 2003 Sub-Cohort					
EJCPS	N	%	Overall	SE	CI95L	CI95U	N	%	Overall	SE	CI95L	CI95U
1	50	1.4	**352.6**	13.46	326.2	379.0	34	1.5	**344.8**	17.57	310.4	379.3
2	175	4.8	**400.8**	5.86	389.3	412.3	133	5.8	**390.7**	7.05	376.9	404.5
3	408	11.2	**445.4**	3.82	437.9	452.9	288	12.4	**432.2**	4.47	423.4	440.9
4	386	10.6	**469.6**	4.34	461.0	478.1	246	10.7	**453.8**	5.12	443.8	463.9
5	619	17.0	**508.2**	3.06	502.2	514.2	387	16.8	**493.1**	3.89	485.5	500.7
6	1104	30.3	**541.3**	2.43	536.6	546.1	698	30.2	**531.0**	3.24	524.6	537.3
7	700	19.2	**578.4**	2.89	572.7	584.0	403	17.4	**565.7**	3.92	558.0	573.4
8	201	5.5	**609.4**	4.48	600.6	618.1	122	5.3	**603.0**	5.79	591.6	614.4
All available	3643	100.0	**518.9**	2.66	513.6	524.1	2312	100.0	**503.2**	3.28	496.8	509.6
Pearson	r = .672, df = 80, p < .001						r = .689, df = 80, p < .001					

Table A4.10. Means, Standard Errors, and 95% Confidence Intervals For the Combined Mathematics Score, and for the Four Mathematics Content Area Subscales, For Each Point on the MCJPS: Scale 1 - Sub-Cohort Taking the Junior Certificate Mathematics Examination in 2003

MJCPS	N	%	Overall	SE	CI95L	CI95U	S&S	SE	CI95L	CI95U	C&R	SE	CI95L	CI95U	U	SE	CI95L	CI95U	Q	SE	CI95L	CI95U
1*	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
2	3	0.1	**316.3**	24.10	269.1	363.5	**281.3**	50.01	183.3	379.4	**278.1**	40.98	197.7	358.4	**319.4**	35.69	249.4	389.4	**307.3**	37.26	234.3	380.3
3	11	0.5	**349.3**	18.17	313.7	384.9	**309.7**	24.58	261.5	357.8	**349.3**	22.54	305.1	393.5	**353.2**	23.25	307.6	398.7	**336.8**	20.68	296.2	377.3
4	57	2.4	**375.5**	8.88	358.1	392.9	**358.3**	10.48	337.8	378.8	**378.0**	10.36	357.7	398.3	**393.9**	9.25	375.8	412.0	**373.9**	8.80	356.7	391.2
5	160	6.9	**394.2**	5.45	383.5	404.8	**369.9**	6.28	357.6	382.2	**392.8**	5.441	382.1	403.4	**408.1**	5.39	397.5	418.7	**388.3**	5.18	378.1	398.4
6	276	11.9	**421.8**	5.66	410.7	432.9	**399.9**	5.29	389.5	410.2	**424.9**	5.132	414.9	435.0	**431.5**	5.18	421.3	441.6	**420.7**	5.75	409.4	431.9
7	386	16.7	**456.3**	3.22	450.0	462.6	**432.5**	4.15	424.4	440.6	**461.2**	3.203	455.0	467.5	**469.6**	3.44	462.8	476.3	**454.6**	3.57	447.6	461.6
8	388	16.8	**484.3**	3.06	478.3	490.3	**458.4**	4.64	449.3	467.5	**485.7**	3.811	478.2	493.1	**499.0**	3.72	491.7	506.3	**482.4**	4.69	473.2	491.6
9	292	12.6	**519.6**	3.60	512.5	526.6	**491.2**	4.09	483.2	499.2	**524.0**	3.871	516.4	531.6	**535.5**	4.16	527.4	543.7	**517.0**	4.00	509.1	524.8
10	283	12.2	**537.4**	3.35	530.8	544.0	**510.5**	4.19	502.3	518.7	**540.6**	3.499	533.8	547.5	**553.4**	3.38	546.8	560.0	**536.3**	3.34	529.8	542.8
11	297	12.8	**565.8**	3.85	558.2	573.3	**539.9**	3.99	532.1	547.8	**570.7**	3.086	564.6	576.7	**583.9**	3.43	577.2	590.6	**566.7**	3.37	560.1	573.3
12	161	6.9	**612.3**	6.01	600.6	624.1	**594.2**	7.63	579.2	609.1	**613.6**	5.446	602.9	624.3	**631.0**	5.57	620.0	641.9	**611.3**	5.43	600.6	621.9
All available	2312	0.1	**492.8**	2.97	486.9	498.6	**468.1**	3.00	462.2	473.9	**495.8**	2.921	490.0	501.5	**507.6**	3.221	501.3	513.9	**491.1**	3.07	485.1	497.1

*No student was awarded Grade F at Foundation Level

Note. 'S&S' = Space & Shape subscale; 'C&R' = Change & Relationships subscale; 'U' = Uncertainty subscale; 'Q' = Quantity subscale.

Table A4.11. Means, Standard Errors, and 95% Confidence Intervals For the Combined Mathematics Score, and for the Four Mathematics Content Area Subscales, For Each Point on the MCJPS: Scale 2 - Sub-Cohort Taking the Junior Certificate Mathematics Examination in 2003

MJCPS	N	%	Overall	SE	CI95L	CI95U	S&S	SE	CI95L	CI95U	C&R	SE	CI95L	CI95U	U	SE	CI95L	CI95U	Q	SE	CI95L	CI95U
1	70	3.0	**369.1**	8.37	352.7	385.5	**347.8**	9.34	329.4	366.1	**369.5**	9.648	350.6	388.4	**384.7**	8.73	367.5	401.8	**365.5**	7.84	350.2	380.9
2	160	6.9	**394.2**	5.45	383.5	404.8	**369.9**	6.28	357.6	382.2	**392.8**	5.441	382.1	403.4	**408.1**	5.39	397.5	418.7	**388.3**	5.18	378.1	398.4
3	276	11.9	**421.8**	5.66	410.7	432.9	**399.9**	5.29	389.5	410.2	**424.9**	5.132	414.9	435.0	**431.5**	5.18	421.3	441.6	**420.7**	5.75	409.4	431.9
4	386	16.7	**456.3**	3.22	450.0	462.6	**432.5**	4.15	424.4	440.6	**461.2**	3.203	455.0	467.5	**469.6**	3.44	462.8	476.3	**454.6**	3.57	447.6	461.6
5	388	16.8	**484.3**	3.06	478.3	490.3	**458.4**	4.64	449.3	467.5	**485.7**	3.811	478.2	493.1	**499.0**	3.72	491.7	506.3	**482.4**	4.69	473.2	491.6
6	292	12.6	**519.6**	3.60	512.5	526.6	**491.2**	4.09	483.2	499.2	**524.0**	3.871	516.4	531.6	**535.5**	4.16	527.4	543.7	**517.0**	4.00	509.1	524.8
7	283	12.2	**537.4**	3.35	530.8	544.0	**510.5**	4.19	502.3	518.7	**540.6**	3.499	533.8	547.5	**553.4**	3.38	546.8	560.0	**536.3**	3.34	529.8	542.8
8	297	12.8	**565.8**	3.85	558.2	573.3	**539.9**	3.99	532.1	547.8	**570.7**	3.086	564.6	576.7	**583.9**	3.43	577.2	590.6	**566.7**	3.37	560.1	573.3
9	161	6.9	**612.3**	6.01	600.6	624.1	**594.2**	7.63	579.2	609.1	**613.6**	5.446	602.9	624.3	**631.0**	5.57	620.0	641.9	**611.3**	5.43	600.6	621.9
All available	2312	0.1	**492.8**	2.97	486.9	498.6	**468.1**	3.00	462.2	473.9	**495.8**	2.921	490.0	501.5	**507.6**	3.221	501.3	513.9	**491.1**	3.07	485.1	497.1

Note. 'S&S' = Space & Shape subscale; 'C&R' = Change & Relationships subscale; 'U' = Uncertainty subscale; 'Q' = Quantity subscale.

Table A4.12. Means, Standard Errors, and 95% Confidence Intervals For the Combined Mathematics Score, and for the Four Mathematics Content Area Subscales, For Each Point on the MCJPS: Scale 3 - Sub-Cohort Taking the Junior Certificate Mathematics Examination in 2003

MJCPS	N	%	Overall	SE	CI95L	CI95U	S&S	SE	CI95L	CI95U	C&R	SE	CI95L	CI95U	U	SE	CI95L	CI95U	Q	SE	CI95L	CI95U
1	230	9.9	**386.5**	4.93	376.8	396.2	**363.2**	5.43	352.5	373.8	**385.7**	4.82	376.2	395.1	**400.9**	5.08	391.0	410.9	**381.3**	4.41	372.7	390.0
2	276	11.9	**421.8**	5.66	410.7	432.9	**399.9**	5.29	389.5	410.2	**424.9**	5.132	414.9	435.0	**431.5**	5.18	421.3	441.6	**420.7**	5.75	409.4	431.9
3	386	16.7	**456.3**	3.22	450.0	462.6	**432.5**	4.15	424.4	440.6	**461.2**	3.203	455.0	467.5	**469.6**	3.44	462.8	476.3	**454.6**	3.57	447.6	461.6
4	388	16.8	**484.3**	3.06	478.3	490.3	**458.4**	4.64	449.3	467.5	**485.7**	3.811	478.2	493.1	**499.0**	3.72	491.7	506.3	**482.4**	4.69	473.2	491.6
5	292	12.6	**519.6**	3.60	512.5	526.6	**491.2**	4.09	483.2	499.2	**524.0**	3.871	516.4	531.6	**535.5**	4.16	527.4	543.7	**517.0**	4.00	509.1	524.8
6	283	12.2	**537.4**	3.35	530.8	544.0	**510.5**	4.19	502.3	518.7	**540.6**	3.499	533.8	547.5	**553.4**	3.38	546.8	560.0	**536.3**	3.34	529.8	542.8
7	297	12.8	**565.8**	3.85	558.2	573.3	**539.9**	3.99	532.1	547.8	**570.7**	3.086	564.6	576.7	**583.9**	3.43	577.2	590.6	**566.7**	3.37	560.1	573.3
8	161	6.9	**612.3**	6.01	600.6	624.1	**594.2**	7.63	579.2	609.1	**613.6**	5.446	602.9	624.3	**631.0**	5.57	620.0	641.9	**611.3**	5.43	600.6	621.9
All available	2312	0.1	**492.8**	2.97	486.9	498.6	**468.1**	3.00	462.2	473.9	**495.8**	2.921	490.0	501.5	**507.6**	3.221	501.3	513.9	**491.1**	3.07	485.1	497.1

Note. 'S&S' = Space & Shape subscale; 'C&R' = Change & Relationships subscale; 'U' = Uncertainty subscale; 'Q' = Quantity subscale.

Table A4.13. Means, Standard Errors, and 95% Confidence Intervals For the Combined Mathematics Score, and for the Four Mathematics Content Area Subscales, For Each Point on the MCJPS: Scale 4 - Sub-Cohort Taking the Junior Certificate Mathematics Examination in 2003

MJCPS	N	%	Overall	SE	CI95L	CI95U	S&S	SE	CI95L	CI95U	C&R	SE	CI95L	CI95U	U	SE	CI95L	CI95U	Q	SE	CI95L	CI95U
1	70	3.0	**369.1**	8.37	352.7	385.5	**347.8**	9.34	329.4	366.1	**369.5**	9.648	350.6	388.4	**384.7**	8.73	367.5	401.8	**365.5**	7.84	350.2	380.9
2	436	18.8	**411.7**	4.09	403.7	419.7	**388.9**	3.77	381.5	396.3	**413.2**	3.636	406.0	420.3	**422.9**	3.77	415.5	430.3	**408.8**	4.03	400.9	416.7
3	386	16.7	**456.3**	3.22	450.0	462.6	**432.5**	4.15	424.4	440.6	**461.2**	3.203	455.0	467.5	**469.6**	3.44	462.8	476.3	**454.6**	3.57	447.6	461.6
4	388	16.8	**484.3**	3.06	478.3	490.3	**458.4**	4.64	449.3	467.5	**485.7**	3.811	478.2	493.1	**499.0**	3.72	491.7	506.3	**482.4**	4.69	473.2	491.6
5	292	12.6	**519.6**	3.60	512.5	526.6	**491.2**	4.09	483.2	499.2	**524.0**	3.871	516.4	531.6	**535.5**	4.16	527.4	543.7	**517.0**	4.00	509.1	524.8
6	283	12.2	**537.4**	3.35	530.8	544.0	**510.5**	4.19	502.3	518.7	**540.6**	3.499	533.8	547.5	**553.4**	3.38	546.8	560.0	**536.3**	3.34	529.8	542.8
7	297	12.8	**565.8**	3.85	558.2	573.3	**539.9**	3.99	532.1	547.8	**570.7**	3.086	564.6	576.7	**583.9**	3.43	577.2	590.6	**566.7**	3.37	560.1	573.3
8	161	6.9	**612.3**	6.01	600.6	624.1	**594.2**	7.63	579.2	609.1	**613.6**	5.446	602.9	624.3	**631.0**	5.57	620.0	641.9	**611.3**	5.43	600.6	621.9
All available	2312	0.1	**492.8**	2.97	486.9	498.6	**468.1**	3.00	462.2	473.9	**495.8**	2.921	490.0	501.5	**507.6**	3.221	501.3	513.9	**491.1**	3.07	485.1	497.1

Note. 'S&S' = Space & Shape subscale; 'C&R' = Change & Relationships subscale; 'U' = Uncertainty subscale; 'Q' = Quantity subscale.

Table A4.14. Means, Standard Errors, and 95% Confidence Intervals For the Combined Mathematics Score, and for the Four Mathematics Content Area Subscales, For Each Point on the MCJPS: Scale 5 - Sub-Cohort Taking the Junior Certificate Mathematics Examination in 2003

MJCPS	N	%	Overall	SE	CI95L	CI95U	S&S	SE	CI95L	CI95U	C&R	SE	CI95L	CI95U	U	SE	CI95L	CI95U	Q	SE	CI95L	CI95U
1	230	9.9	386.5	4.9	376.8	396.2	**363.2**	5.43	352.5	373.8	**385.7**	4.82	376.2	395.1	**400.9**	5.08	391.0	410.9	**381.3**	4.41	372.7	390.0
2	276	12.0	421.8	5.7	410.7	432.9	**399.9**	5.29	389.5	410.2	**424.9**	5.132	414.9	435.0	**431.5**	5.18	421.3	441.6	**420.7**	5.75	409.4	431.9
3	377	16.3	455.8	3.2	449.4	462.1	**432.0**	4.31	423.5	440.4	**460.8**	3.29	454.3	467.2	**468.9**	3.49	462.1	475.8	**454.1**	3.65	447.0	461.3
4	353	15.3	482.9	3.2	476.5	489.3	**457.3**	5.02	447.4	467.1	**484.4**	4.189	476.2	492.7	**498.4**	4.08	490.4	506.4	**480.8**	5.32	470.4	491.3
5	100	4.3	513.5	6.0	501.7	525.3	**486.9**	7.27	472.6	501.1	**515.3**	6.927	501.7	528.9	**526.5**	6.67	513.4	539.6	**506.4**	6.00	494.6	518.1
6	8	0.4	480.9	16.9	447.8	514.0	**454.9**	21.22	413.3	496.5	**482.6**	15.69	451.8	513.3	**499.1**	22.00	456.0	542.2	**476.2**	17.93	441.1	511.4
7	35	1.5	499.1	11.4	476.8	521.3	**470.2**	12.37	445.9	494.4	**498.0**	13.64	471.2	524.7	**505.2**	15.01	475.8	534.6	**498.7**	12.56	474.1	523.3
8	191	8.3	522.7	4.8	513.4	532.1	**493.4**	5.18	483.3	503.6	**528.5**	4.403	519.9	537.2	**540.3**	5.34	529.8	550.7	**522.5**	4.91	512.9	532.2
9	283	12.2	537.4	3.4	530.8	544.0	**510.5**	4.19	502.3	518.7	**540.6**	3.499	533.8	547.5	**553.4**	3.38	546.8	560.0	**536.3**	3.34	529.8	542.8
10	297	12.9	565.8	3.9	558.2	573.3	**539.9**	3.99	532.1	547.8	**570.7**	3.086	564.6	576.7	**583.9**	3.43	577.2	590.6	**566.7**	3.37	560.1	573.3
11	161	7.0	612.3	6.0	600.6	624.1	**594.2**	7.63	579.2	609.1	**613.6**	5.446	602.9	624.3	**631.0**	5.57	620.0	641.9	**611.3**	5.43	600.6	621.9
All available	2312	0.1	**492.8**	2.97	486.9	498.6	**468.1**	3.00	462.2	473.9	**495.8**	2.921	490.0	501.5	**507.6**	3.221	501.3	513.9	**491.1**	3.07	485.1	497.1

Note. 'S&S' = Space & Shape subscale; 'C&R' = Change & Relationships subscale; 'U' = Uncertainty subscale; 'Q' = Quantity subscale.

Table A4.15. Means, Standard Errors, and 95% Confidence Intervals For the Combined Mathematics Score, and for the Four Mathematics Content Area Subscales, For Each Point on the MCJPS: Scale 6 - Sub-Cohort Taking the Junior Certificate Mathematics Examination in 2003

MJCPS	N	%	Overall	SE	CI95L	CI95U	S&S	SE	CI95L	CI95U	C&R	SE	CI95L	CI95U	U	SE	CI95L	CI95U	Q	SE	CI95L	CI95U
1	230	9.9	**386.5**	4.93	376.8	396.2	**363.2**	5.43	352.5	373.8	**385.7**	4.82	376.2	395.1	**400.9**	5.08	391.0	410.9	**381.3**	4.41	372.7	390.0
2	276	12.0	**421.8**	5.66	410.7	432.9	**399.9**	5.29	389.5	410.2	**424.9**	5.132	414.9	435.0	**431.5**	5.18	421.3	441.6	**420.7**	5.75	409.4	431.9
3	377	16.3	**455.8**	3.23	449.4	462.1	**432.0**	4.31	423.5	440.4	**460.8**	3.29	454.3	467.2	**468.9**	3.49	462.1	475.8	**454.1**	3.65	447.0	461.3
4	353	15.3	**482.9**	3.25	476.5	489.3	**457.3**	5.02	447.4	467.1	**484.4**	4.189	476.2	492.7	**498.4**	4.08	490.4	506.4	**480.8**	5.32	470.4	491.3
5	109	4.7	**511.0**	5.56	500.1	521.9	**484.4**	6.77	471.1	497.6	**512.8**	6.321	500.4	525.2	**524.4**	5.94	512.7	536.0	**504.1**	5.65	493.0	515.1
6	35	1.5	**499.1**	11.35	476.8	521.3	**470.2**	12.37	445.9	494.4	**498.0**	13.64	471.2	524.7	**505.2**	15.01	475.8	534.6	**498.7**	12.56	474.1	523.3
7	191	8.3	**522.7**	4.77	513.4	532.1	**493.4**	5.18	483.3	503.6	**528.5**	4.403	519.9	537.2	**540.3**	5.34	529.8	550.7	**522.5**	4.91	512.9	532.2
8	283	12.2	**537.4**	3.35	530.8	544.0	**510.5**	4.19	502.3	518.7	**540.6**	3.499	533.8	547.5	**553.4**	3.38	546.8	560.0	**536.3**	3.34	529.8	542.8
9	297	12.9	**565.8**	3.85	558.2	573.3	**539.9**	3.99	532.1	547.8	**570.7**	3.086	564.6	576.7	**583.9**	3.43	577.2	590.6	**566.7**	3.372	560.1	573.3
10	161	7.0	**612.3**	6.01	600.6	624.1	**594.2**	7.63	579.2	609.1	**613.6**	5.446	602.9	624.3	**631.0**	5.57	620.0	641.9	**611.3**	5.434	600.6	621.9
All available	2312	0.1	**492.8**	2.97	486.9	498.6	**468.1**	3.00	462.2	473.9	**495.8**	2.921	490.0	501.5	**507.6**	3.221	501.3	513.9	**491.1**	3.07	485.1	497.1

Note. 'S&S' = Space & Shape subscale; 'C&R' = Change & Relationships subscale; 'U' = Uncertainty subscale; 'Q' = Quantity subscale.

Table A4.16. Means, Standard Errors, and 95% Confidence Intervals For the Combined Mathematics Score, and for the Four Mathematics Content Area Subscales, For Each Point on the MCJPS: Scale 7 - Sub-Cohort Taking the Junior Certificate Mathematics Examination in 2003

MJCPS	N	%	Overall	SE	CI95L	CI95U	S&S	SE	CI95L	CI95U	C&R	SE	CI95L	CI95U	U	SE	CI95L	CI95U	Q	SE	CI95L	CI95U
1	230	9.9	**386.5**	4.93	376.8	396.2	**363.2**	5.43	352.5	373.8	**385.7**	4.82	376.2	395.1	**400.9**	5.08	391.0	410.9	**381.3**	4.41	372.7	390.0
2	276	11.9	**421.8**	5.66	410.7	432.9	**399.9**	5.29	389.5	410.2	**424.9**	5.132	414.9	435.0	**431.5**	5.18	421.3	441.6	**420.7**	5.75	409.4	431.9
3	377	16.3	**455.8**	3.23	449.4	462.1	**432.0**	4.31	423.5	440.4	**460.8**	3.29	454.3	467.2	**468.9**	3.49	462.1	475.8	**454.1**	3.65	447.0	461.3
4	361	15.6	**482.9**	3.15	476.7	489.0	**457.2**	4.96	447.5	466.9	**484.4**	4.176	476.2	492.6	**498.4**	4.00	490.5	506.2	**480.7**	5.22	470.5	491.0
5	135	5.8	**509.8**	5.68	498.6	520.9	**482.6**	6.83	469.2	496.0	**510.8**	6.152	498.8	522.9	**521.0**	6.73	507.8	534.2	**504.4**	5.43	493.8	515.1
6	191	8.3	**522.7**	4.77	513.4	532.1	**493.4**	5.18	483.3	503.6	**528.5**	4.403	519.9	537.2	**540.3**	5.34	529.8	550.7	**522.5**	4.91	512.9	532.2
7	283	12.2	**537.4**	3.35	530.8	544.0	**510.5**	4.19	502.3	518.7	**540.6**	3.499	533.8	547.5	**553.4**	3.38	546.8	560.0	**536.3**	3.34	529.8	542.8
8	297	12.8	**565.8**	3.85	558.2	573.3	**539.9**	3.99	532.1	547.8	**570.7**	3.086	564.6	576.7	**583.9**	3.43	577.2	590.6	**566.7**	3.372	560.1	573.3
9	161	6.9	**612.3**	6.01	600.6	624.1	**594.2**	7.63	579.2	609.1	**613.6**	5.446	602.9	624.3	**631.0**	5.57	620.0	641.9	**611.3**	5.434	600.6	621.9
All available	2312	0.1	**492.8**	2.97	486.9	498.6	**468.1**	3.00	462.2	473.9	**495.8**	2.921	490.0	501.5	**507.6**	3.221	501.3	513.9	**491.1**	3.07	485.1	497.1

Note. 'S&S' = Space & Shape subscale; 'C&R' = Change & Relationships subscale; 'U' = Uncertainty subscale; 'Q' = Quantity subscale.

Table A4.17. Pearson Correlations Between Eight MJCPS Scales and PISA 2003 Mathematics: Combined Scale and Subscales - Sub-Cohort Taking the Junior Certificate Mathematics Examination in 2003

	MJCPS Scale							
PISA Scale	*Scale 1*	*Scale 2*	*Scale 3*	*Scale 4*	*Scale 5*	*Scale 6*	*Scale 7*	*Scale 8*
Combined Scale	.766	.766	.764	.760	.742	.751	.760	.766
Space and Shape	.682	.681	.680	.676	.658	.667	.675	.684
Change and Relationships	.750	.749	.747	.742	.726	.735	.743	.748
Uncertainty	.755	.755	.754	.751	.732	.741	.750	.755
Quantity	.745	.745	.743	.738	.722	.731	.739	.744

Correlations are all significant (p<.001).

Table A4.18. Means, Standard Errors, and 95% Confidence Intervals For the Combined Mathematics Scale, PISA 2000 Sample (All Students Attempting PISA 2000 Mathematics, and the Sub-Sample Attempting the Junior Certificate Examination in 2000)

	PISA 2000 - All Participating & Attempting PISA Mathematics						*PISA 2003 - JCE 2000 Sub-Cohort & Attempting PISA Mathematics*					
MJCPS	*N*	*%*	*Overall*	*SE*	*CI95L*	*CI95U*	*N*	*%*	*Overall*	*SE*	*CI95L*	*CI95U*
1	189	9.5	**392.1**	5.49	381.4	402.9	142	11.0	**385.9**	5.97	374.2	397.6
2	248	12.4	**448.5**	5.09	438.5	458.5	179	13.9	**445.7**	5.51	434.9	456.5
3	376	18.8	**480.3**	4.08	472.3	488.3	234	18.1	**471.9**	4.84	462.4	481.4
4	339	17.0	**504.2**	3.63	497.1	511.3	213	16.5	**495.8**	5.04	485.9	505.7
5	284	14.2	**539.9**	4.14	531.8	548.0	189	14.6	**532.5**	5.02	522.7	542.3
6	236	11.8	**551.5**	4.10	543.4	559.5	142	11.0	**546.0**	4.47	537.2	554.7
7	226	11.3	**579.7**	4.20	571.5	588.0	134	10.4	**572.6**	4.95	562.9	582.3
8	99	5.0	**611.9**	6.32	599.5	624.3	58	4.5	**605.0**	7.83	589.6	620.3
All available	1997	100.0	**506.7**	2.59	501.6	511.8	1290	100.0	**496.2**	3.05	490.2	502.1
Pearson	r = .700, df = 80, p < .001						r = .698, df = 80, p < .001					

Table A4.19. Means, Standard Errors, and 95% Confidence Intervals For the Space & Shape Mathematics Subscale, PISA 2000 Sample (All Students Attempting PISA 2000 Mathematics, and the Sub-Sample Attempting the Junior Certificate Examination in 2000)

	PISA 2000 - All Participating & Attempting PISA Mathematics						PISA 2003 - JCE 2000 Sub-Cohort & Attempting PISA Mathematics					
MJCPS	*N*	*%*	*S&S*	*SE*	*CI95L*	*CI95U*	*N*	*%*	*S&S*	*SE*	*CI95L*	*CI95U*
1	189	9.5	**396.4**	8.43	379.9	413.0	142	11.0	**396.4**	9.23	378.3	414.5
2	248	12.4	**433.3**	8.32	417.0	449.6	179	13.9	**433.6**	9.87	414.2	452.9
3	376	18.8	**454.4**	5.19	444.3	464.6	234	18.1	**451.8**	7.03	438.0	465.6
4	339	17.0	**472.9**	5.36	462.4	483.4	213	16.5	**464.3**	8.09	448.5	480.2
5	284	14.2	**501.6**	7.27	487.4	515.9	189	14.6	**495.7**	9.30	477.5	513.9
6	236	11.8	**507.4**	7.41	492.9	521.9	142	11.0	**509.0**	8.66	492.0	526.0
7	226	11.3	**531.5**	6.95	517.9	545.1	134	10.4	**527.4**	8.43	510.9	543.9
8	99	5.0	**566.5**	11.81	543.3	589.6	58	4.5	**566.1**	13.09	540.4	591.8
All available	1997	100.0	**476.7**	3.30	470.2	483.1	1290	100.0	**470.9**	4.26	462.6	479.3
Pearson	r = .450, df = 80, p < .001						r = .452, df = 80, p < .001					

Table A4.20. Means, Standard Errors, and 95% Confidence Intervals For the Change & Relationships Mathematics Subscale, PISA 2000 Sample (All Students Attempting PISA 2000 Mathematics, and the Sub-Sample Attempting the Junior Certificate Examination in 2000)

	PISA 2000 - All Participating & Attempting PISA Mathematics						PISA 2003 - JCE 2000 Sub-Cohort & Attempting PISA Mathematics					
MJCPS	*N*	*%*	*C&R*	*SE*	*CI95L*	*CI95U*	*N*	*%*	*C&R*	*SE*	*CI95L*	*CI95U*
1	189	9.5	**424.1**	7.49	409.4	438.8	142	11.0	**423.9**	7.97	408.3	439.5
2	248	12.4	**460.8**	6.77	447.5	474.1	179	13.9	**459.1**	6.89	445.6	472.6
3	376	18.8	**483.9**	4.08	475.8	491.9	234	18.1	**480.2**	5.71	469.0	491.4
4	339	17.0	**499.2**	4.69	490.0	508.4	213	16.5	**491.3**	6.36	478.8	503.8
5	284	14.2	**525.1**	4.57	516.1	534.1	189	14.6	**518.0**	5.87	506.5	529.6
6	236	11.8	**538.0**	7.49	523.3	552.7	142	11.0	**536.9**	9.32	518.7	555.2
7	226	11.3	**559.8**	6.13	547.8	571.8	134	10.4	**553.8**	7.23	539.6	567.9
8	99	5.0	**588.4**	10.18	568.4	608.3	58	4.5	**583.8**	12.19	559.9	607.7
All available	1997	100.0	**503.9**	2.69	498.7	509.2	1290	100.0	**496.9**	3.14	490.8	503.1
Pearson	r = .505, df = 80, p < .001						r = .495, df = 80, p < .001					

APPENDIX 5: ADDITIONAL TABLES

Table A5.1. *Description of Items Comprising ESCS and School-Level Composites in Models of Students in PISA 2000, PISA 2003 and TIMSS 1995/1996*

Composite	*PISA 2000 and 2003*	*TIMSS 1995*
Collected at the school level		
School educational resources	In your school, how much is the learning of third years hindered by: lack of instructional material (e.g., texbooks), not enough computers for instruction, lack of instructional materials in the library, lack of multimedia resources for instruction, inadequate science laboratory equipment, inadequate facilities for the fine arts. Scale: not at all, very little, to some extent, a lot.	Is your school's capacity to provide instruction affected by a shortage or inadequacy of any of the following: Instructional materials (e.g., textbooks), computers for mathematics instruction, computer software for mathematics instruction, calculators for mathematics instruction, library materials relevant to mathematics instruction, audio-visual resources for mathematics instruction, science laboratory equipment and materials. Scale: none, a little, some, a lot.
Student behaviour	In your school, is the learning of third years hindered by: student absenteeism, disruption of classes by students, students skipping classes, students lacking respect for teachers, the use of alcohol or illegal drugs, students intimidating or bullying other students. Scale: not at all, very little, to some extent,a lot.	About how often does the school administration or staff have to deal with the following behaviours among second year students? Absenteeism, skipping classes, classroom disturbance, intimidation or verbal abuse of teachers or staff, intimidation or verbal abuse of other students, alcohol use/possession, drug use/possession. Scale: rarely, monthly, weekly, daily.
School autonomy	In your school, who has the main responsibility for: appointing teachers, dismissing teachers, establising teachers' starting salaries, determining teachers' salary increases, formulating the school budget, deciding on budget allocations within the school, establishing student disciplinary policies, establishing student assessment policies, approving students for admittance to the school, choosing which textbooks are used, determining course content, deciding which courses are offered. Response categories: option to tick not a school responsibility, appointed board, principal, department head, teachers. Ticked responses for not a school responsibility used to construct the composite.	With regard to your school, who has the primary responsibility for each of the following activities? Appointing teachers, establishing disciplinary policies, formulating the school budget, determining which textbooks are used, establishing homework policies, determining teacher salaries, determining course content, deciding which courses are offered. Response categories: option to tick not a school responsibility, appointed board, principal, department head, teachers. Ticked responses for not a school responsibility used to construct the composite.

Table A5.1. *Continued.*

Composite	*PISA 2000 and 2003*	*TIMSS 1995*
Teacher participation	In your school, who has the main responsibility for: appointing teachers, dismissing teachers, establising teachers' starting salaries, determining teachers' salary increases, formulating the school budget, deciding on budget allocations within the school, establishing student disciplinary policies, establishing student assessment policies, approving students for admittance to the school, choosing which textbooks are used, determining course content, deciding which courses are offered. Response categories: option to tick not a school responsibility, appointed board, principal, department head, teachers. Ticked responses for teachers used to construct the composite.	With regard to your school, who has the primary responsibility for each of the following activities? Appointing teachers, establishing disciplinary policies, formulating the school budget, determining which textbooks are used, establishing homework policies, determining teacher salaries, determining course content, deciding which courses are offered. Response categories: option to tick not a school responsibility, appointed board, principal, department head, teachers. Ticked responses for teachers used to construct the composite.
School building quality	In your school, how much is the learning of third years hindered by: poor condition of buildings, poor heating, cooling and/or lighting systems, lack of instructional space (e.g., classrooms). Scale: not at all, very little, to some extent, a lot.	Is your school's capacity to provide instruction affected by a shortage or inadequacy of any of the following: school buildings and grounds, heating/cooling and lighting systems, instructional space (e.g., classrooms). Scale: none, a little, some, a lot.
Collected at the student level		
School disciplinary climate	How often to these things happen in your English/mathematics classes? The teacher has to wait a long time for students to settle down, students cannot work well, students don't listen to what the teacher says, students don't start working for a long time after the lesson begins, there is noise and disorder, at the start of the class, more than five minutes is spent doing nothing. Scale: never, some lessons, most lessons, every lesson.	In my mathematics class… students often neglect their school work, students are orderly and quiet during class, students do exactly as the teacher says. Scale: strongly agree, agree, disagree, strongly disagree.
ESCS	Composite based on higher of parents' occupation, higher of parents' education, and a range of educational and material home possessions (desk for study, own room, quiet place to study, computer for school work, educational software, link to the Internet, own calculator, classic literature, books for school work, dictionary, works of art, dishwasher. The variable also included a binary variable indicating >100 books in the home. The relative weights given to each of these components in the construction of ESCS is not known.	Composite based on higher of parents' education, books in the home (in its logarithmic form), access to a calculator, desk, dictionary, encyclopaedia, phone, dishwasher, microwave, tumble dryer, and second bathroom. A weighted average, assigning twice the weight to parental education and books in the home, was constructed and then re-scaled to have a student mean of 1.0 and standard deviation of 0.0.

Table A5.2. Factor Loadings for TIMSS ESCS Component Educational Possessions

Item	*Loading*
Calculator	.584
Desk	.546
Dictionary	.603
Encyclopaedia	.581

Percent of variance explained: 33.5%.

Table A5.3. Factor Loadings for TIMSS ESCS Component Material Possessions

Item	*Loading*
Phone	.633
Microwave	.597
Dishwasher	.692
Tumble dryer	.557
Second bathroom	.682

Percent of variance explained: 40.3%.

Table A5.4. Factor Loadings for TIMSS Disciplinary Climate

Item	*Loading*
Neglect work	-.650
Orderly and quiet	.845
As teacher says	.843

Percent of variance explained: 61.5%.

Table A5.5. Number of Responses Not Ticked for the Eight School Autonomy Variables

	Frequency	*% All Cases*	*% All Available*
Four	1	0.8	0.8
Five	9	6.8	7.1
Six	60	45.5	47.2
Seven	51	38.6	40.2
Eight	6	4.5	4.7
All Available	127	96.2	100
Missing	5	3.8	
Total	132	100	

Note. School autonomy was transformed to have a mean of 0.0 and standard deviation of 1.0 using the following: compute sch. auton = (variable as shown in Table - 5.92125984252)/0.4473253709567

Table A5.6. Number of Responses Ticked for the Eight Teacher Participation Variables

	Frequency	*% All Cases*	*% All Available*
None	19	14.4	15.0
One	21	15.9	16.5
Two	44	33.3	34.6
Three	28	21.2	22.0
Four	14	10.6	11.0
Five	1	0.8	0.8
All Available	127	96.2	100
Missing	5	3.8	
Total	132	100	

Note. Teacher participation was transformed to have a mean of 0.0 and standard deviation of 1.0 using the following: compute tch. parti = (variable as shown in Table - 2.0000000000)/1.227980662688

Table A5.7. Factor Loadings for TIMSS School Building Quality

Item	*Loading*
Building	.882
Heat/cool/light	.735
Space	.875

Percent of variance explained: 69.4%.

Table A5.8. Factor Loadings for TIMSS Educational Resources

Item	*Loading*
Instructional material	.566
Computer hardware (maths)	.711
Computer software (maths)	.784
Calculators	.634
Library materials (maths)	.838
Audiovisual equipment (maths)	.787
Science lab. equipment	.626

Percent of variance explained: 50.8%.

Table A5.9. Factor Loadings for TIMSS Student Behaviour

Item	*Loading*
Absenteeism	.626
Skipping class	.686
Class disturbance	.544
Bullying students	.724
Intimidating teachers	.655
Alcohol use	.651
Drug use	.618

Percent of variance explained: 41.7%.