Using and Applying International Survey Data on Mathematics and Science Education

Thomas G Macintyre

Doctor of Philosophy
The University of Edinburgh
2014

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Thomas G. Macintyre

Acknowledgments

I would like to acknowledge and thank my supervisors Prof Lindsay Paterson and Prof David Raffe for their patience, advice and guidance in supporting me to complete this study. I appreciated the feedback and comments that made such sense on re-visiting original drafts and for assisting me to scale down the original scope of this thesis.

I would also thank the Applied Quantitative Methods Network (AQMeN) and Prof Susan McVie in the University of Edinburgh for the programme of events and training sessions that were organised for Ph.D. candidates. I found the courses, talks and Ph.D. training particularly interesting and supportive in providing a network of like-minded students in what can otherwise be an isolated existence. Prof McVie's feedback at progression stage was also much appreciated.

Another source of inspiration and support came from the Centre for Multilevel Modelling (CMM) at the University of Bristol. Thanks to that team for their LEMMA (Learning Environment for Multilevel Methodology and Applications) online course that was extremely clear and informative for a self-study exercise, and of course for all the developments they have put into the MLwiN software that I have relied upon throughout my study period.

I would like to acknowledge the funding I received through an ESRC/Scottish Government PhD Studentship and also my employer, The University of Edinburgh, for granting leave of absence with return to substantive post on completion of the studentship.

Finally I cannot overstate my appreciation of the support and encouragement I have received from my family. Celia and Zoe in particular have put up with a lot of late nights with lights on and holidays without me as I pushed to finish this thesis. Thanks for believing in me– I'm looking forward to repaying you both as best I can.

I'd like to dedicate this thesis to my mother who has always taken a keen interest in my education and in memory of my late father who never saw it completed but was always supportive of my progress. I would also like to dedicate this thesis in memory of my niece Esme Morris Macintyre whose life was cut so short she never had the opportunity to fulfil her dreams or education.

Abstract

There were two purposes set out in this study, first to identify the principal associations with educational performance of Scottish students as reported in the 2007 wave of the Trends in International Mathematics and Science Study ($TIMSS_{2007}$), and second to evaluate methods of data analysis where sample surveys use plausible value (PV) methodology.

Four sets of data were used for the secondary analysis of $TIMSS_{2007}$, with students' responses to cognitive items and questionnaire data emanating from two stages (G4 and G8) that each addressed two disciplines (mathematics and science). Explanatory models for each stage and discipline were analysed using hierarchical linear modelling techniques to accommodate the cluster sample design of the survey. Guided by existing literature in STEM education the study examined elements of students' learning experiences that fell within a social constructivist theory of learning to ascertain whether the empirical data supported current claims on effective practice. A number of control variables were included in the analyses, some well-established constructs and others derived from background questionnaires. Overall, the results showed that selected background characteristics were consistently related to mathematics and science achievement. The strength of association varied across disciplines with scientific achievement showing a stronger association with home resources, and although girls were generally associated with lower achievement scores, that gender association was strongest in G4 mathematics achievement. The findings suggest there is limited support for current claims in respect of a reform agenda that privileges discussion and collaborative group work. Other policy initiatives on assessment for learning and using technologies in class are not supported in the data, with either no evidence of association or a significant negative effect in the models of mathematics and science achievement. Aspects of practical work and scientific enquiry are positively associated with G4 science achievement, with particular credence given to 'doing' and 'watching' experiments or investigations, but there is no association with achievement scores at G8 for any of planning, watching or conducting experiments. This latter finding provides empirical evidence of difference across stages on an aspect of practice that is heavily debated.

The primary method of analysis utilised a four-level structure, with PV as the unit of analysis. Substantive findings were compared with alternative methods: first making the dependent variable an average of the five PVs; second using one PV as the response variable; and third computing statistics from all five PVs and merging results using Rubin's Rules for combining multiple imputations. The findings from these alternative analyses suggest the primary multilevel method underestimates standard errors in the model in the same way as witnessed for the average of PVs. This leads to the conclusion that the only valid route to analysing imputed data is through Rubin's method of combining results from all five PVs.

Contents

1 Introduction

The role of assessment in education has featured highly in recent years as policy initiatives have emphasised the varied dimensions of assessment under the broad agenda of 'Assessment is for Learning' (AifL). Three major types of assessment can be used to support student achievement and to influence educational policy within AifL: assessment *for* learning that focuses on formative feedback and diagnostic assessment; assessment *as* learning that allows students to use feedback to further their own learning at a metacognitive level; and assessment *of* learning that often takes the form of summative assessment used to report on progress and for accountability purposes. At a national level, the Scottish Survey of Achievement (SSA) provided data and information against the 'assessment *of* learning' dimension of AifL to the profession and policy makers. The survey data provided comment on a national picture and as such aimed to provide *'*a robust national monitoring system that provides reliable information about overall standards and trends*'* (Scottish Government, 2009b). This national measure has now been replaced by a Scottish Survey of Literacy and Numeracy (SSLN). This sample survey of learners' attainment in Scottish Primary and Secondary schools is designed to monitor national performance in literacy and numeracy, reporting on each discipline in alternate years for learners in Primary 4, Primary 7 and Secondary 2[1], having started with Numeracy in 2011. The Scottish Government uses findings from the survey to facilitate improvements in learning, teaching and assessment at classroom level, through publication of professional learning resources for practitioners. There are limitations however in the way those professional learning resources are compiled because learning and teaching practices reported by teachers and learners through questionnaires are not explicitly linked to learners' performances, a matter that has prompted this exploration of survey data and alternative methods of analyses. At an international level, the Scottish Government was committed to participating in two surveys that were relevant to mathematics and science education. First the Programme for International Student Assessment (PISA) that was launched by the Organisation for Economic Co-operation and Development (OECD) in 1997 to evaluate education systems by testing the skills and knowledge of 15 year old learners, and second the Trends in International Mathematics and Science Study (TIMSS) that is organised and coordinated by International Association for the Evaluation of Educational Achievement (IEA). This second survey has a focus on student achievement and factors related to it. Both PISA and TIMSS provide the Scottish Government with international benchmarks and data from comparator countries to

[1] Scottish Secondary 2 is equivalent to Year 9 in England and Wales where learners are aged 13-14 years

complement national data used to monitor standards and achievements within mathematics and science education in Scottish schools. Subsequent to starting this research the Scottish Government decided to restrict its commitment to International surveys and now only participates in the PISA study. The last participation in TIMSS was in 2007, so the Scottish Government no longer has any comparator benchmark data for Primary or lower Secondary stages of its education system.

There is currently limited reporting of findings from any of these three sources of data and I believe more could be gained by both policy makers and the teaching profession if the data were analysed using advanced quantitative techniques to identify the principal associations with educational performance. In particular, my study seeks to identify strengths and weakness in learning to directly support the planning of future teaching programmes and to target resources accordingly. A challenge will be to identify ways of supporting schools and authorities with their improvement planning for numeracy, mathematics and science. A further challenge will be to get the most out of the national and international survey data and to make those data available in a suitable and useable format for local analyses and action. If public-funded survey data can be utilised in a more productive way, by policy makers and practitioners, I would argue there is scope for an overall reduction in individual student assessment at a time when summative assessment and assessment *of* learning can appear unnecessarily to be dominating activity in schools.

The attraction of developing statistical techniques and skills has been influential in pursuing this project. Using and understanding the underlying principles of statistical analyses in social sciences is clearly of the utmost importance for educational research, the teaching profession and policy development. As such, I would like to contribute to those developments through effective communication of methods and practice, supporting the users as opposed to the producers of research data, drawing on my expertise in learning and teaching to maximise the messages and to encourage fuller and wider use of survey data in support of policy and practice in mathematics and science education.

1.1 Background information on surveys informing educational policy

The Scottish Government has invested in survey data collection and analysis for many years, to monitor national educational standards and experiences. This investment has been at a national and international level.

Nationally, an annual survey of education has been in place since 1983, with the Assessment of Achievement Programme (AAP) providing data on national performance levels based on a representative cluster sample of pupils, stratified by local authority and school size. This rolling programme of monitoring standards initially covered three areas of the curricula, English language, mathematics and science, but it was extended to include data on Social Subjects from 2002 onwards. The AAP was replaced by the SSA in 2005, continuing the rolling programme of curricular coverage and providing a national picture of performance levels across different stages (P3, P5, P7 and S2). On the international stage, Scotland has participated in the TIMSS since its inception in 1995. $TIMSS_{2007}$ is the fourth in a series of comparative international surveys of mathematics and science achievement administered on a four-yearly cycle. The survey is targeted at students across two stages of schooling, first Grade 4 (G4) where students are aged nine to ten years and second Grade 8 (G8) with students aged thirteen to fourteen years. The survey design is based on stage of learning rather than age. The best agreed match for Scotland's participation in $TIMSS_{2007}$ equates to Primary 5 and Secondary 2 where the average ages were 9.8 years and 13.7 years respectively. This sample of students provides a convenient overlap with two of the stages sampled in SSA and SSLN.

These national and international surveys provide feedback to education authorities and teachers, and directly inform policy analysts and curriculum developers with a view to improving learning and teaching. However, there is limited evidence of the latter being progressed explicitly on the basis of survey findings, hence the desire to develop an analytical framework that can maximise the use of these and subsequent surveys. The evidence is assembled and headline findings are published (Mullis *et al.*, 2008; Horne *et al.*, 2008) but deeper key findings do not appear to reach those that matter. This study will consider how best to redress that negative observation by focusing on practitioners and policy analysts' needs and desires, to ensure they can readily access what can be taken from the surveys to meet their requirements.

Given the historical differences within the educational systems of the United Kingdom it has been important for Scotland to be represented on the international stage as separate from the UK when participating in international surveys of education (PISA &

TIMSS). Scotland's interests are represented at the planning stages with a representative from Scottish Government taking a front seat at negotiations. This is seen as very important in terms of maintaining a high profile on the international arena. For instance there are two representatives from the UK on the PISA planning group, one from Scotland and a representative from the Department for Education representing England, Wales and Northern Ireland. This position is not something that other comparable regions can lay claim to, as articulated by one of Grek's respondents:

> I think we probably started to play more active part in the governing board of PISA, that would be fair. So that in itself has been quite interesting. The fact that we have a place nobody else does so for example the Spanish representative, there will be one representative sitting at the table representing Spain and three of four people sitting behind as observers representing Catalan etc. We're the only country that managed that and that caused some confusion (Grek *et al.*, 2009a: 17)

In the current political climate, with the Scottish National Party (SNP) leading the administration and government in Scotland, it can be argued that there are additional and wider political reasons for Scotland to be represented as an individual state in these international surveys. Grek suggests that one way the SNP administration can highlight differences in policy between Scotland and the UK (Labour government at time of writing) is to shift its point of reference from England and to highlight Scotland's similarities to small continental European countries participating in PISA (Grek *et al.*, 2009b). In a similar way, with Scotland reported as a separate entity in TIMSS it makes the Scottish educational system visible to a wide audience and ensures a seat on the international education policy stage.

1.2 Research questions

The PhD study seeks to report on the use and application of survey data on mathematics and science education. Two purposes are set out in this study, first to identify the principal associations with educational performance of Scottish students as reported in the 2007 wave of the TIMSS, and second to evaluate methods of data analysis where sample surveys use plausible value methodology. My decision to focus on the $TIMSS_{2007}$ data rested on the fact that this survey spanned Primary and Secondary stages of education as well as addressing mathematics and science cognitive domains. The structure of data as disseminated by IEA was also influential in choosing $TIMSS_{2007}$ as the focus of study. The nature of the data, with its hierarchical structure and linkage within each strand of the sample survey provided an opportunity to investigate alternative methods of analysing multilevel

data built on plausible value methodology. In particular, a new multilevel plausible value method of analysis is proposed and evaluated; findings are compared with those generated through traditional methodology using Rubin's rules for imputed data. The analyses and suggested methodologies are targeted to support practitioner and policy communities to use the data and to apply the findings from large scale survey data in education. An overriding driver is to make full and effective use of sample survey data in order to reduce assessment *of* learning for accountability purposes in schools. Specifically, the study aims to address the following research questions:

1) What are the principal associations with educational performance in mathematics and science for Scottish students in $TIMSS_{2007}$?

   (a) Which background characteristics of G4 and G8 students are strongly associated with educational performance in mathematics and science?

   (b) Having controlled for principal background characteristics, which pedagogical practices are strongly associated with educational performance in mathematics and science?

   (c) What differences, if any, can be reported across disciplines within each stage?

   (d) What differences, if any can be reported across stages within each discipline?

2) How can publicly-funded national and international survey data be used to fuller effect, to influence policy and practice?

   (a) Can an analytical framework be developed as a basis for the secondary analysis of survey data relevant to practice and policy development in mathematics and scientific literacies?

   (b) How does the proposed multilevel plausible value method compare with the standard practice of using Rubin's method for the analysis of imputed data?

1.2.1 Rationale and background to research questions

1. What are the principal associations with educational performance in mathematics and science for Scottish students in $TIMSS_{2007}$?

This aspect will be theory driven from the evidence base of effective teaching of mathematics and science. The literature on learning and teaching, and research on *scientific*

literacy and *working mathematically* will be used to develop a conceptual framework for the analysis of achievement data in $TIMSS_{2007}$. Sources will include the national reports from Her Majesty's Inspectorate of Education (HMIE) as well as comparable publications within the UK and from an international perspective. The promotion of teaching approaches and learner activities to focus on conceptual understanding of curricular topics is a feature of recent research and evaluation of learning and teaching in mathematics and science education. A starting point for the analysis will be to ascertain whether there is evidence of those practices being adopted, and what relationship those experiences have with performance in the cognitive domains of the surveys. A primary focus will therefore be on students' experiences and teachers' practices as reported through questionnaires. These instruments cover in-school as well as out-of-school experiences, the latter offering insight to social and economic factors that influence learners' educational attainment and that can be used as control variables in models of educational achievement. The pedagogical experiences documented through the questionnaire responses will be organised in accordance with a conceptual framework from the theory. IEA include derived factors in the published data set, but an exercise for this study and subsequent investigation will be to use principal components analysis (PCA) as a method for further data reduction to support the analyses. Regression analyses will be used to interpret the data and to explain the variance within the survey by examining differences in educational achievement that are associated with background and pedagogical variables. Models of achievement scores will take account of the hierarchical structure of the data, making use of multilevel modelling techniques for each discipline and stage of study. The unit of analysis in the proposed method will be plausible value, where there are five plausible values assigned to each student in the sample. There will be four data sets to accommodate the two stages and two disciplines. Analyses will identify pedagogical practices that provide a significant explanation of variance in the models of achievement for each data set.

2. *How can publicly-funded national and international survey data be used to fuller effect, to influence policy and practice?*

Recent literature on summative assessment acknowledges the concern that assessment can dominate school experiences, unduly influencing learning and teaching as practitioners are drawn to 'teach to the test' (Mansell *et al.*, 2009). This is as a result of learners' assessments being used for multiple purposes, serving as a proxy measure to monitor wider elements of the system including teachers, schools and authorities; as well as providing feedback to the learner. A concern is that learners may lose out, with their long-term educational needs not being met as the priority shifts to making the system look good.

If practitioners do 'teach to the test' by cramming information, drilling learners in techniques at the expense of conceptual understanding, and overlooking the needs of the individual, then only short-term benefits will be seen (Mansell *et al.*, 2009).

I would argue there is scope for learning to be firmly placed at the heart of school experience if an overall reduction in summative assessment for monitoring of standards and quality can be secured. This could be achieved through more intelligent use of survey data. This study will explore whether sample survey data can be analysed to more productively service the needs of stakeholders. If sample survey data can be shown to address issues of standards and quality, and to provide sufficient evidence of practice for policy and practitioner communities then there is every possibility of reducing the overall assessment burden on students.

Regression analysis and Multilevel Modelling (MLM) techniques will enable better interrogation of the data to provide clear messages for stakeholder communities. A desirable output from this study will be a robust and user-friendly analytical framework that is grounded in the theories of learning and teaching for mathematics and science. Such a framework could support the analysis of existing surveys as well serving as a guide in the development of future studies. Analyses will focus on evaluation of policy initiatives and associated practices that research literature has promoted, providing evidence of impact on achievement that can subsequently be used influence policy and practice in schools. The proposed multilevel plausible value method of analysis will also be evaluated to ascertain whether findings based on this new method provide comparable substantive conclusions as would be identified through traditionally accepted methods i.e. using Rubin's method for analysis of multiple imputations. If findings are comparable then this proposed method would open opportunities for a more direct analysis than is currently recommended. The present lack of in-depth analyses of survey data suggests the need for such a framework and review of statistical methods, with a prime focus on generating a structure and method of analysis that will make cross-survey comparisons a more likely and entirely feasible option for educational researchers.

1.3 Research outline

1.3.1 Research design

I propose three strands that will influence the project as it evolves over the study period. First a review of literature, providing an opportunity for articulation of issues from a local, national and international perspective that will guide analysis of survey data; second an exploratory analysis of survey data, building on published summaries and investigating associations between practice and achievement; and third exploring methodological issues pertaining to survey design, focusing on treatments for missing data, methods used to quantify educational achievement, and development of advanced quantitative methods for secondary analyses of the data based on plausible value methodology.

Literature

The emphasis on practice and experience in both the curricular frameworks (Education Scotland, 2009b) and survey data questionnaires links with Dewey's practical epistemology where the central concept in his theory of knowing is the notion of experience (Dewey, 1916). For Dewey, learning experiences underpin thinking and the training of thinking in education. A lot is lost for the learner if that element of experience is by-passed for whatever reason. Dewey cites situations where techniques are taught without 'wasting time' on the early experiences (trial and error approaches; playing with the problem or data), with the expectation that thinking and learning will result. Not so, *'*thinking *is* the method of intelligent learning*'* and *'*[the] initial stage of that developing experience which is called thinking is *experience'*. The worst case scenario is where there is an element of pretence:

> … the problem of the pupil is not how to meet the requirements of school life, but how to seem to meet them – or, how to come near enough to meeting them to slide along without an undue amount of friction. (Dewey cited in Cahn, 2009:444)

The emphasis on *experiences* is at the heart of recent curricular reform in Scotland (Education Scotland, 2009b) where the documentation is primarily presented in terms of 'experiences and outcomes'. Similarly in these surveys, a wide range of questions are asked of participants to ascertain the type of experiences that are self-reported, permitting analysis of how those experiences relate to cognitive attainment. Are learners being given adequate opportunity to think and learn through meaningful experiences as envisaged by Dewey? Alternatively, is there evidence of learning without genuine experience but rather a reliance on the acquisition of skills without thinking i.e. a form of rote learning that is highlighted in

a negative light by Dewey:

> And skill obtained apart from thinking is not connected with any sense of the purposes for which it is to be used. It consequently leaves a man at the mercy of routine habits and of the authoritative control of others. (Dewey, cited in Cahn, 2009:443)

The role of an education professional in this process, policy advisor or classroom practitioner, is not to translate general rules into particular and specific actions but more to be placed in a position of 'knowing' and to make smarter use of research findings and survey data in response to the philosophical questions posed on learning and teaching. The more that can be extracted from the survey data, in terms of useable and useful knowledge of educational practice, the greater the benefits for policy advisors and practitioners.

In summary, the literature on effective learning and teaching, and research on *working mathematically* and *scientific literacy* will be used to guide the analysis of survey data from $TIMSS_{2007}$.

Survey Data

A prime aim of this study is to explore the international survey data on mathematics and science in order to identify the key factors that are associated with educational achievement. TIMSS was selected because of its potential links to the national programmes of assessment used for summarising and reporting on learning in Scottish Primary and lower Secondary stages of school, namely SSA and its successor the SSLN. The international and national surveys overlap in stages and therefore inform the mathematics and science communities of how Scottish learners are performing over time as well as facilitating comparison to their international counterparts. The survey intervals, participating cohorts and subject foci are summarised in Table 1.3-1.

The Scottish Government produces a summary report on the annual SSA with a more detailed technical volume published separately (Scottish Government, 2009b). Over recent surveys, neither of those publications has offered a deep analysis of the curricular subjects, contexts or situations included in the data collection. The reports are primarily limited to summary statistics, stopping short of analysis *per se*, as the publications focus on score aggregation and headline findings from questionnaire responses.

Table 1.3-1: Historical national and international survey data

Survey	Interval	Cohorts	Focus
SSA	Annual	P3, P5, P7, S2	Numeracy and other core skills on an annual basis plus subject specific focus: Science (2007) Mathematics (2008)
TIMSS	4-year	P5, S2	Mathematics and Science (2007)

The IEA publish the findings from TIMSS (Mullis *et al.*, 2008) and make the entire data sets available for participating countries to undertake further analyses. A number of country-specific perspectives are published or commissioned by governments. For example *How Finns Learn Mathematics and Science* (Pehkonen *et al.*, 2007) has been published to reflect on and to offer an explanation for Finland's success in PISA (2000 & 2003) and TIMSS (1999); a comprehensive report on England's achievement in $TIMSS_{2007}$ was published as a national report that focused on English learners' high performance in the study and compared performance with those of other European neighbours and economic competitors (Sturman *et al.*, 2008). In Scotland, a brief report on the highlights of Scotland's results from $TIMSS_{2007}$ was published (Horne *et al.*, 2008), but nothing of a more substantial nature has emerged, leaving the analysis of Scottish data at a very basic level alongside the headline findings published by Mullis *et al.* (2008). There is clearly an opportunity for researchers to explore the underlying features of learners' profiles and to identify factors that influence educational achievements; a gap that this study seeks to fill. Comparisons can be made with learners from high-achieving countries, and also with learners from countries that have similar characteristics to Scotland in terms of their population and other organisational structures for mathematics and science education, but the focus in this thesis is on the within-country data and methods of analysis.

Analysis

A major emphasis in the current political climate and policy development in education is the concept of evidenced-based policy (EBP). The EBP movement has been developed on the back of successful experiences in the field of medicine, where experimental research and quantitative methods are often viewed as the *gold standard*. In social science research there are difficulties in deploying that *standard* through undertaking experiments and randomised controlled trials (RCTs), but that doesn't detract from the principles of EBP or diminish the desire for social scientists to have strong links between

research and policy. The social sciences and education research operate with a range of systematic methods and approaches, an evidence base that is probably more diverse than that witnessed in the sciences, but clearly accepted as providing valid evidence without necessarily meeting the 'gold standard' referred to above.

In education research there is a softening of the EBP argument, where policy development is viewed as an on-going process that is strongly influenced by past experiences and contemporary social and political contexts. Research evidence is seen as contributing to the policy development but is not the strict *basis* of policy as implied by EBP, but rather informing the development (Bridges *et al.*, 2009). This leads Nutley *et al.* (2009) to favour 'evidence influenced', or even just 'evidence aware' to reflect a more realistic view of what can be achieved. That said, they continue to use the accepted shorthand of EBP or evidence-based policy and practice (EBPP) to acknowledge acceptance of the wider principles of the movement that underpin the policy process.

With political pressure to come up with *quick fix* solutions there is a tendency to present somewhat limited and instrumental outcomes from research that focus on 'what works'. The 'what works' agenda of EBP is insufficient in the case of education, because judgment in education is about what is educationally desirable. This is a value judgment that goes beyond what can be accessed through a technical assessment of what is possible and one that must take into consideration the social, economic and political contexts of the educational setting. Biesta (2007) and Sanderson (2003) argue for the need to go beyond such a technical approach if true progress and benefits are to be accrued from research. This position is further supported by Levin (2013) in his discussion of knowledge mobilisation:

> It is clear that research never determines practice in a rote fashion; professionals are always making judgements about which research to apply, in what way, under what circumstances. Far from deskilling teachers, the use of research findings can actually increase professional skill and discretion. (Levin, 2013:8)

Whilst there is support for a utilitarian focus and for useful and useable outcomes, if not policy development, there are concerns over the limitations that a 'what works' agenda can impose (Solesbury and ESRC, 2001). So first there is the question about what it should work for, and who has a say in what makes an 'effective school', a 'good teacher' or 'positive educational experiences'. The deeper philosophical questions that underpin the 'what works' movement need to be considered before research can inform policy and practice; and who decides on such philosophical questions becomes more complex when extending research to embrace diverse international perspectives and priorities in education.

In the survey data for this study, the socio-cultural settings and experiences of participants (learners and educators) are accessed through the qualitative questionnaires within the survey data. This information provides insight to the learning contexts and provides evidence beyond the technical assessment of what works, at a national or international level. Such information must be taken into account to avoid determining future actions based on what works for what may be an entirely different context.

Second, researchers must acknowledge that teacher development is a moral activity that involves value judgements over what is educationally desirable. There is no causal explanation or indeed any general explanation that can be claimed as an outcome of the research but rather more tentative suggestions and interpretations that are considered by education professionals. Explanation is at the heart of any analysis, and given the complex socio-cultural contexts within education, it is expected there will be multiple explanations for policy advisors, curriculum development and teacher development. Multiple explanations and interpretations of survey data are therefore anticipated:

> If research is to inform teaching, then evidence of many different kinds has to shape the views (or 'hypotheses') of the teacher (Oancea and Pring, 2009:22)

Evidence from the international survey will be analysed with that in mind, reflecting on the educational experiences of learners and educators within Scotland. My prime focus in this study is a within-country analysis, to note the learning and teaching experiences and their association with achievement, rather than looking to international comparisons as the primary purpose of analysing international survey data. The experiences that are likely to form the basis of such analyses are outlined in the review of literature.

The primary method of analysis takes account of the hierarchical structure of TIMSS data, where samples of students are nested within classes and schools. The IEA employed a two-stage stratified cluster sample design. Scottish schools were stratified by size, urbanisation, and a school deprivation index before the first stage selection of schools was sampled with probability proportional to size. At the second stage, one or more intact classes of students from the target grade were sampled; up to four classrooms were sampled per Primary school and up to three classrooms per Secondary school. A second design feature of $TIMSS_{2007}$ concerns the matrix sampling of cognitive items. The survey design dictates respondents only complete a sub-set of cognitive tasks with other items noted as missing; in effect these items are missing at random (MAR), a feature that will play a significant part in the methodological approaches to the analysis. The achievement scores reported for each student are known as *plausible values,* with five separate plausible values

generated for each student using item response theory (IRT) and scaling distributions that describe the population. Fuller details on plausible value methodology and the imputation of achievement scores are provided in Chapter 3; suffice to say at this stage, that this second design feature of $TIMSS_{2007}$, which results in using plausible value methodology, requires particular methods of analysis to be deployed.

The resultant sample comprises nested data that cannot be analysed using traditional ordinary least squares analysis; the analyses must take into account the clustered nature of the data and the fact that variance within clusters is reduced relative to variance in the same sample size of independent cases. A multilevel structure and method of analysis can take cluster variance into account, analysing the hierarchical data at various levels of the model. A four-level multilevel structure is adopted in this particular analysis with school, teacher, and student-levels modelling achievement data, working with plausible value score as the unit of analysis. A number of variants on this primary model are explored in the study, with analyses providing insights to encourage practitioners to engage with research and to analyse their own data with confidence.

Communication of findings to stakeholders is deemed key to EBP and will be a central feature of this project, recognising that:

> Research will not have impact unless potential users are interested enough to look for it and able to make use of what they find. It is vital to develop the capacity of users to find, understand and use research. (Levin, 2004: 15)

Engaging the end-user is clearly important for any system benefits to be accrued. Research, policy and professional practice need to be connected and the metaphor of bridges between 'research and policy', and 'research and practice' is often cited (Young *et al.*, 2002; Nutley *et al.*, 2003). The difficulty with such an image however, is that it leaves the parties on either side of a divide where ideally we want interaction and engagement across those divisions, perhaps building 'causeways' rather than bridges, to maximize the connection and impact of research on both policy and practice. Although Levin's earlier work cited above emphasised the individual user and intrinsic motivation, a later piece acknowledges the role of organisations and institutions and the need for researchers to engage with institutions, local authorities, managers and school leaders as a way of forging connections between research, policy and practice under the auspices of knowledge mobilisation:

> the use of research is fundamentally a social and organizational process. Whether people are interested in, pay attention to and make use of research evidence depends much more on their organizational setting and social relations than it does on their individual background or dispositions. (Levin, 2013:10)

The proposal to develop an analytical framework for the secondary analysis of survey data is designed to make research findings accessible to practitioners and to provide a means of using and applying past, present and future survey data. Solesbury and ESRC (2001) criticise social researchers for focusing on new primary data at the expense of a wealth of past research already accumulated but mostly ignored in terms of analysis. Secondary analysis of data is recognised as an established method and is currently receiving strong support and attention through specific funding from research councils (Economic & Social Research Council, 2013) and should therefore be encouraged. The current study seeks to maximize the use of existing survey data.

This study reports on student data in $TIMSS_{2007}$. Future work can take forward the analyses of teacher data in a similar way, reporting on associations between achievement and teacher education, qualifications and professional development. Further extensions can take the resultant findings on Scottish mathematics and science education data, and relate findings to comparable cohorts and experiences as reported through SSA and SSLN surveys. The cross-survey analyses will highlight common features and trends that are evidenced in the data, potentially linking the findings from $TIMSS_{2007}$ to National surveys:

i. SSA_{2007} – common cohorts in numeracy and science at P5 and S2 (G4 & G8)

ii. SSA_{2008} – matching stages of P5 and S2 for mathematics and numeracy

1.4 Data issues

The proposed methods for data collection and analyses in relation to each research question are summarised in Table 1.4-1.

Table 1.4-1: Data collection methods and analysis

Research Question	Data collection – methods & analysis
1. What are the principal associations with educational performance in mathematics and science for Scottish students in $TIMSS_{2007}$?	Scrutiny of survey documentation and reports from IEA. Literature review of country specific analyses, in the first instance drawing on Pehkonen (2007) & Sturman (2008). Literature review of Mathematics and Science Education – Learning & Teaching. Principal Component Analysis used in an exploratory fashion to reduce data and to identify suitable latent (derived) variables. EDA of $TIMSS_{2007}$ – Four sets of data used for the secondary analysis, with students' responses to cognitive items and questionnaire data emanating from two stages (G4 and G8) that each addressed two disciplines (mathematics and science); weighted data used for EDA. Generate long data set for each set of data – duplicating records to have 'plausible value' of unit of analysis. Primary analyses – multilevel analyses using 4-level structure to identify principal associations with achievement data (PV)
2. How can publicly-funded national and international survey data be used to fuller effect, to influence policy and practice? In particular, can an analytical framework be developed as a basis for the secondary analysis of survey data relevant to practice and policy development in mathematics and scientific literacies, with a view to reducing the amount of testing and summative assessment in Scottish schools?	A critical review of methods used in survey. Replicate analyses as necessary on the same data sets to confirm initial findings, using alternative approaches by taking response variable as: i. Average of five plausible values ii. A single plausible value iii. All five plausible values analysed with results combined using Rubin's Rules for multiple imputations. Evaluate alternative approaches in comparison to substantive findings reported using Rubin's method, for future researchers to draw conclusions and recommendations on research practice where sample surveys use plausible value methodology.

2. Mathematics and Science Education – Learning and Teaching

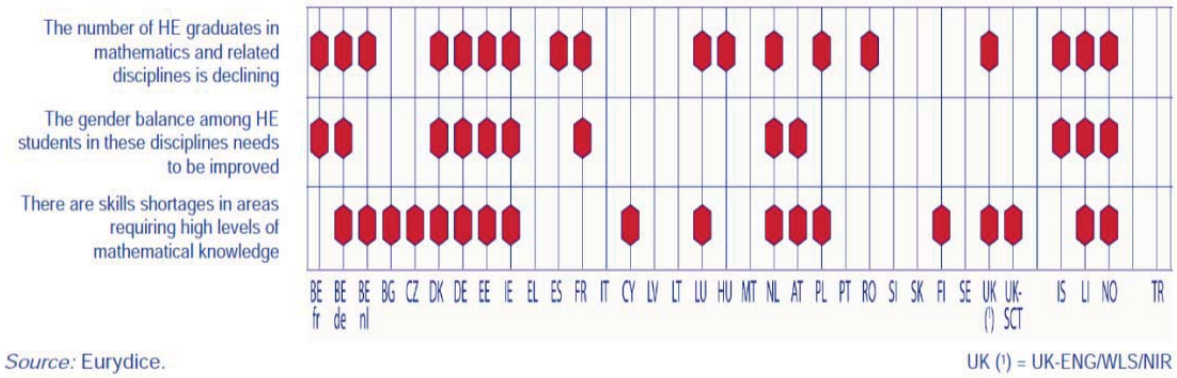
National and international governments acknowledge the economic importance of Science, Technology, Engineering and Mathematics (STEM) skills, and the need for a strong supply of suitably qualified people to be the world-class scientists of tomorrow (Roberts, 2002; Gluckman, 2011; Eurydice Network, 2011; Science and Engineering Education Advisory Group (SEEAG), 2012). The UK government commissioned the Roberts' review to inform its strategy to secure a supply of high quality scientists and engineers in the knowledge that ambitions for economic growth in the 'knowledge economy' would otherwise be seriously constrained. Recommendations for action were made across all levels of education and extended to employment of scientists and engineers working in research and development. The Roberts' review recommended subject-specific training in mathematics and the physical sciences for primary teachers, covering initial teacher education and continuing professional development. Similarly, Roberts recommended subject-specific training and development for secondary school science practitioners, and for those recruiting teachers to pay more attention to their areas of specialism, acknowledging the need for an adequate supply of teachers able to teach individual sciences, particularly physics and chemistry. Another aspect that Roberts highlighted as influential in encouraging students to study those subjects at higher levels, concerned the condition and standard of school science laboratories; the recommendation being for these to be 'representative of the world of science', in order to inspire and motivate students to continue their studies. That subset of recommendations outlines how the education system is viewed as a driving force in helping to secure a strong supply of people with science and engineering skills to satisfy the UK Government's agenda to position the nation as one of the world's best.

In Scotland, the devolved administration's economic strategy (Scottish Government, 2007) places a similar focus on science and engineering with Energy and Medical/Life Sciences identified as priority sectors in the overall strategy. Scotland has an historic record of being a *Science Nation* and the Scottish Government wishes to build on that strength in its economic strategy. In international terms, the research base in Scotland is identified as a significant strength, ranking first among 26 comparator nations (relative to Gross Domestic Product (GDP)) that collectively account for 95 per cent of the world's research in the area of science engineering and technology (Scottish Government, 2009b). It is important for Scotland to realise that economic benefit from areas of world class research and to capitalise on key opportunities in associated industries. SEEAG (2012) highlights the

importance of young people continuing with STEM subjects to sustain the supply of scientists and engineers to maintain that legacy and to drive forward innovative technologies within the Energy and Life Science sectors. Scotland needs to build the skills base that can deliver the anticipated growth in output from the energy sector as a whole, including clean fossil fuels, carbon capture and storage, and renewable energies. The Scottish Government claims Scotland has a competitive advantage in renewable energy, with offshore and marine energy projected to create 35000 direct jobs and to contribute an extra £11 billion in Gross Value Added (GVA) by 2020 (Scottish Government, 2009b); a significant impact on the economic model, where GDP = GVA + taxes on products – subsidies on products. The Scottish Government priorities have changed slightly in the updated economic strategy (Scottish Government, 2011c) but there remains a commitment to helping business and creating jobs in growth sectors that include *Energy*, *Life Sciences* (including medical technologies and pharmaceutical services) and *Creative Industries*, all of whom need to recruit scientifically literate people with aptitude and skills in science and technology (Scottish Government, 2009b).

Similar arguments and national strategies can be witnessed across Europe and further afield. A European Commission report on mathematics education in Europe (Eurydice, 2011) outlines some common concerns, challenges and suggested actions from 31 countries and regions surveyed. One of the driving issues for many regions relates to the perceived economic impact of being unable to sustain a supply of personnel to satisfy an increasing shortage of a highly qualified workforce in mathematics, science and technology (MST). Several countries already indicate a shortage of highly qualified personnel in mathematics and related fields, with 18 nations indicating a skills shortage as shown in lower portion of Figure 2-1. UK and UK-Scotland are included in that number.

Figure 2-1: Policy concerns related to skills shortages and the take-up of mathematics and related disciplines in higher education, 2010/11 (Eurydice Network, 2011: 107)



Source: Eurydice.

UK (1) = UK-ENG/WLS/NIR

Among other actions, strategies are being put in place to increase motivation in

mathematics learning, often within a broad policy for promoting the learning and teaching of MST. For instance in Finland, support is provided at all levels through developing activities for learners including MST camps, as well as training and workshops for teachers. A lot of this support is channelled through the *LUMA* Centre, an organisation coordinated by the University of Helsinki that supports links between schools, universities, business and industry.

In Norway, a strategy for strengthening MST has been put in place under the umbrella name of *Science for the Future*. The Norwegian Ministry of Education and Research acknowledges that high competencies in MST are required to meet the challenges of tomorrow, and sufficient numbers of people with confidence and understanding in those disciplines are needed to comprehend those challenges, and to act accordingly. The Norwegian government wants to create a knowledge society in which everyone can participate, noting that 'not everyone shall specialize in MST, but all require knowledge in Mathematics and Science topics' (Norwegian Ministry of Education and Research, 2010: 5). As with Scotland, the Norwegian government projects the need for an increased pool of candidates for employment in the energy sector and other technology-driven business that largely depend on scientific expertise. Norway projects a very similar development to that predicted by the Scottish Government in terms of renewable energy, capture and storage of carbon dioxide (CO_2), and the associated impact on employment that comes with meeting the European Union (EU) targets on renewable energy. These jobs will require employees to be well-educated with a particular demand for engineers and skilled workers with expertise in MST. The Norwegian strategy therefore places a greater emphasis on mathematics and science throughout education to increase skills and motivation in sciences among students and teachers to increase recruitment to MST subjects.

Not only is there need to have a supply of scientists, engineers and mathematicians to develop and exploit new technologies, there is also a call for highly qualified individuals to educate and inspire those future generations of STEM-qualified personnel. The economic benefits of satisfying such a supply of skills has been argued above but there is also a call for the wider populace to be scientifically literate. For example, New Zealand's response to the increasing demand of an economy based on knowledge and innovation has been to focus attention on an educational strategy (Gluckman, 2011). This strategy was informed by a report entitled *Inspired by Science* (Bull *et al.*, 2010) that distinguished two strands for a future curriculum: pre-professional education and citizen-informed objectives. The pre-professional education, traditional studies in science, lead directly to further and higher education that builds on those STEM skills to satisfy the employment demands for STEM-

qualified personnel. Whereas the citizen-informed objectives address the need to educate all learners in three distinct ways – first with practical knowledge of how things work, to understand and appreciate the technological world of work and leisure (utilitarian purpose); second to gain knowledge in the scientific process and to have sufficient scientific literacy skills to participate in society, able to make informed decisions and to contribute to science-related decisions in communities (democratic / citizenship purpose); and third to develop scientific thinking skills that emulate the cognitive processes of the classical scientists, valuing the processes and structures of scientific knowledge and placing scientific thinking as an explicit goal for learners (cultural / intellectual purpose). It is those citizen-informed objectives that highlight the importance of science for a wider population in the coming decades. Bull *et al.* (2010) acknowledge a shift in emphasis from traditional studies in science, including specific knowledge development in physics, chemistry, biology and mathematics, to a preparation for life in a technological world that is relevant to all society. Gluckman (2011) argues that any future education system has to satisfy both strands: the need to ensure enough young people exit school with aptitude and skills for transition to scientific and technological professions; and to ensure a 'healthy democracy' is created where the population at large is able to participate in discussions and debates of a scientific nature, acting in an informed capacity as a result of experiences in the education system and beyond. Osborne and Hennessy (2003) report similar findings for the future of science education in the UK, highlighting four categories and purposes of science education very much as captured above, using the labels of *economic, utilitarian, cultural,* and *democratic*.

So what does STEM education entail? What are the major factors that influence STEM teachers' practice? In addressing these questions I propose to focus on a subset of STEM education, paying particular attention to the influences that affect the learning of mathematics and science in primary and lower secondary schools. My focus on learning places children's experiences at the centre of the study, with learning theories considered in relation to teaching styles and school contexts. First I will discuss learning theories that have a particular bearing on mathematics and science education and students' understanding of those disciplines in schools. Second I expand on the various shades of constructivism as a grand theory of learning within mathematics and science education. Third I consider argumentation as an important component of learning that strengthens students' confidence in knowledge acquisition and understanding, and teachers' facilitation of constructivist learning theories; and finally, I review the purposes and perceived benefits of practical activities in support of mathematics and science learning. This chapter concludes with an outline of claims that will be investigated through the empirical data, linking the operational

variables with experiences and practices reviewed as prevalent within mathematics and science education.

2.1. Key aspects of learning and teaching in study

In recent decades there has been a shift in emphasis on learning experiences with a focus on active learning strategies in support of constructivist theories on learning. When accounting for that shift in emphasis, current literature in mathematics and science education distinguishes between learning experiences through reference to 'traditional' and so-called 'reform' teaching (Watson, 2008). The former reflects a heavily teacher-oriented experience in the classroom, often associated with teacher's exposition and worked examples that are followed by students working on textbook style exercises, often functioning on their own and by and large devoid of any active learning or self-regulation of practice (Niemi, 2002). The latter reference to reform teaching, acknowledges the pedagogical changes that have been instigated through various curricular reforms in education, including for instance major curricular developments in Scotland such as 5-14 Guidelines (Scottish Office Education Department, 1991), Standard Grade (Scottish Qualification Authority (SQA), 2001) and Curriculum for Excellence (Education Scotland, 2009a); and in literature from the USA related to educational reforms that followed the publication of their curriculum and professional *Standards* for school mathematics (National Council of Teachers of Mathematics (NCTM), 1989, 1991) and subsequent publication of *Common Core Standards* (National Governors Association (NGA), 2010). Indeed the publication of those curriculum and professional Standards is what brought the 'traditional' and 'reform' references to the fore, through what became known as the 'The Math Wars' (Schoenfeld, 2004). Traditionalists clashed with reformers, at times in heated public debate, over the proposed changes in curricula and recommended pedagogies as presented in the Standards documents.

The vision set out by NCTM's publications focused on new goals for society, and students in particular, with five general goals for all students K-12, namely:

> (1) that they learn to value mathematics, (2) that they become confident in their ability to do mathematics, (3) that they become mathematical problem solvers, (4) that they learn to communicate mathematically, and (5) that they learn to reason mathematically (NCTM, 1989:5).

Traditionalists were concerned over the shift in pedagogical focus from memorisation and practice, to one where process skills were emphasised over content coverage, with an underlying assumption of learning being an active process with varied forms of instruction including: 'appropriate project work; group and individual assignments; discussion between teacher and students and among students; practice on mathematical

methods; and exposition by the teacher' (NCTM, 1989:10). The controversial publication of content lists that identified topics to receive increased and decreased attention in the reformed curriculum fuelled the debate, with traditionalists fearing the new curriculum was superficial in coverage and undermined classical mathematical values of rigour and depth of treatment. Not dissimilar debates were pursued in UK around the same time, voiced through The London Mathematical Society (LMS, 1995) in response to concerns that recent changes in school mathematics do not provide the necessary foundations to maintain the quantity and quality of mathematically competent school leavers. They argued this was something that seriously disadvantaged those who wish to continue their mathematical studies beyond school level. Included in their concerns and suggested causes were recent pedagogical changes resulting from recommendations in the Cockcroft report (1982: para 243[1]), associated advice from HMI (1985), and wording from National Curriculum documentation (School Curriculum and Assessment Authority (SCAA), 1994):

> In recent years English school mathematics has seen a marked shift of emphasis, introducing a number of time-consuming activities (investigations, problem-solving, data surveys, etc.) at the expense of 'core' technique. In practice, many of these activities are poorly focused; moreover, inappropriate insistence on working within a context uses precious time and can often obscure the underlying mathematics... (LMS, 1995: 10)

> [W]e have also seen implicit 'advice' ... that teachers should reduce their emphasis on, and expectations concerning, technical fluency. This trend has often been explicitly linked to the assertion that 'process is at least as important as technique'. Such advice has too often failed to recognise that to gain a genuine understanding of any process it is necessary first to achieve a robust technical fluency with the relevant content ... (LMS, 1995; 10)

Many of those concerns in both US and UK were voiced in the absence of real data. It was only many years later at the turn of the 21st Century when large-scale data could evaluate the impact of curricular changes. Schoenfeld (2004) reports that when the evidence was considered, it was unambiguously in favour of the reform movement (ARC Center, 2003), but as far as the politicised 'Math Wars' were concerned, data did not matter. The Alternatives for Rebuilding Curricula (ARC, 2003) findings show that the average mathematics scores of students in the reform schools were significantly higher than the

[1] The Cockcroft report (1982) presents a very similar list of suggested experiences as referenced in the later publication by NCTM (1989). The most often cited paragraph is number 243: Mathematics teaching at all levels should include opportunities for

- exposition by the teacher;
- discussion between teacher and pupils and between pupils themselves;
- appropriate practical work;
- consolidation and practice of fundamental skills and routines;
- problem solving, including the application of mathematics to everyday situations;
- investigational work.

average scores of students in their matched comparison schools; of 34 comparisons across five state-grade combinations, 28 were in favour of the reform students, six show no statistically significant difference, and none favoured the comparison students. Fuson *et al.* (2000) present evidence in support of the reform movement, defending the criticism of promoting 'fuzzy mathematics', by showing that students who followed a programme that embraced the ideas within NCTM's *Standards* (1989, 1991) were not disadvantaged. Indeed, performing equally well as students using traditional approaches on standard vertical symbolic addition and subtraction calculations and outperforming other groups when addressing wider aspects of mathematics that were regarded as conceptually demanding. Schoenfeld (2002) supports the same view insofar as basic skills are concerned, noting there are no significant performance differences between students' learning experiences (traditional- or reform-based) but that on tests of conceptual understanding those students who learn through reform-based practice 'consistently outperform students who learn from traditional curricula by a wide margin' (Schoenfeld, 2002:16). In the same paper Schoenfeld raises some wider implications of reform-based practice as it impacts on equity and civil rights, noting that that 'traditional performance gaps between majority students and poor or underrepresented minorities are diminished, though not eliminated.' (Ibid, 2002: 14).

A review of learning theories and associated experiences in contemporary mathematics and science education presents a reference point for subsequent analyses. Within this review of learning theories, specific consideration is given to the place of malleable process factors, including: memorisation of facts, processes, formulae and procedures; active learning; practical activities and opportunities to watch / design-plan-conduct experiments; the role of talk, discussion and argumentation; collaborative learning; review of homework and other aspects of assessment; and relating learning to authentic contexts and to students' daily life.

2.2. Learning theories

There has been considerable debate over recent decades about how learners develop their mathematical and scientific thinking. A major change in learning and teaching that is embedded within the 'reform' agenda discussed above, has been the introduction and acceptance of constructivist principles to classroom learning. But such practices have not been without criticism and expression of concern, particularly from some within the mathematics and science communities (Millar, 1989; Solomon, 1994; Phillips, 1995; Fosnot, 1996; Airasian and Walsh, 1997; Geelan, 1997; Simpson, 2002; Ford, 2010; Taber, 2011), who have difficulties reconciling 'individual construction' with the fact that mathematics and science as disciplinary knowledge is, at any particular point in time, almost entirely a 'body of consensually agreed knowledge' (Millar, 1989).

Constructivism is an epistemology, a philosophical explanation about the nature of knowledge and a theory about how people learn, with a clear focus on classroom experiences (Airasian and Walsh, 1997; Fosnot, 1996). Geelan (1997) and Simpson (2002) both contend that a constructivist epistemology to teaching and learning is far superior to transmissive epistemologies and teaching approaches, such as those associated with behaviourist theory. The strength of a constructivist theory is in its focus on concept development as opposed to explaining learning as a system of behavioural responses to stimuli (including reinforcement, external motivation, and practice). As a theory of learning it describes how connected structures and deep conceptual understanding come about, through active and student-centred processes, including use of authentic activities and experiences that engage learners. The fundamental premise for constructivism is that people create knowledge from the interaction between their existing knowledge and any new ideas or experiences they encounter. Airasian and Walsh (1997) highlight the motivational aspect of pursuing a constructivist agenda as an additional benefit, with reference to 'lighting the flame' rather than 'filling the bucket' of students' heads with facts; the latter often associated with a transmission model of learning where the student is viewed as an empty vessel or blank slate (*tabula rasa*) and learning is regarded as a change in behaviour following exposure to external stimuli. That strong emphasis on observed behaviour and external assessment ignores the contribution to learning that can be provided through self-evaluation or personal monitoring of understanding and progress; the self-regulation of practice that is embedded within constructivist learning (Fosnot, 1996; Niemi, 2002).

Constructivism as a world view of learning was developed in response to a broad review of existing theories. In any constructivist learning model, students are placed at the heart of the process with their prior knowledge used as the start point. The focus on the

learner and learning process places greater emphasis on 'thinking about learning' rather than a subject or topic that is taught. Very close attention is paid to the process of knowledge construction, paying due attention to learner's mental processes and structures as they actively construct and evaluate meanings from experiences.

How those meanings are developed, both individually and collectively, and how the connections and mental schema are structured to support cognitive development has evolved to produce a range of constructivist frameworks that influence education. The core ideas are not new with the works of Whitehead and Dewey underpinning constructivist practices. Whitehead (1929) stressed the importance of making connections within learning and of keeping learning alive, viewing it as a continuous and active process like a living organism. Students who do not make connections in their learning, but who rather adopt Whitehead's portrayal of learning as the 'passive reception of disconnected ideas', appear unable to use or apply their knowledge when challenged to do so in new contexts. Whitehead defined *inert ideas* as those 'ideas that are merely received into the mind without being utilized, or tested, or thrown into fresh combinations', and advocated that education was only worthwhile if it is purposefully used by learners; if left unused there is no reason in learning. This can be witnessed in students who study intently for exams, potentially achieve a high grade, but who promptly forget everything they learned as soon as they move beyond the class into employment or further study. This inability to transfer knowledge to new applications is consequentially very limiting for students when it comes to problem solving and investigative work that has held a prominent position in mathematics and science curricula since the latter part of the 20th Century. Such deficiencies in students' strengths of connections in and transferability of knowledge raise questions over whether they have *understood* what has been *learned*.

Dewey's contribution to the development of constructivism rests in his emphasis on the prominence given to experiences in education. His practical epistemology, where learning experiences underpin thinking, contends that knowledge only emerges from those situations in which leaners have to draw them out of meaningful experiences (Dewey, 1916). There are many facets of meaningful experience that align themselves with Dewey's constructivism. High on any list would be the social and contextual nature of experience, in that learning and understanding is most effectively supported when developed and discussed with other learners and educators in classrooms and wider communities of practice (Wenger, 2006), where they construct knowledge together. Students cannot learn by means of rote memorisation alone; in Dewey's vision of learning they can only learn through direct experience and anything less will be inferior. Millar supports this view, emphasising that

any knowledge must be reconstructed by the learner in the learning process and that:

> We cannot teach a body of knowledge by direct transmission; the learner is always involved in reconstructing the meaning personally (Millar, 1989: 592).

If that element of experience is by-passed for whatever reason, the learner misses out on potential conceptual and cognitive development that combines practice with theory. Dewey cites situations where techniques are taught without 'wasting time' on early learning experiences (trial and error approaches; playing with the problem or data), with the expectation that thinking and learning will result. However, Dewey argues against any such expectation as he firmly establishes thinking as an essential aspect of intelligent learning. He goes on to stress that any student who does not equate thinking with learning cannot make the necessary connections to understand concepts being studied, where any

> skill obtained apart from thinking is not connected with any sense of the purposes for which it is to be used. It consequently leaves a man at the mercy of routine habits and of the authoritative control of others. (Dewey cited in Cahn, 2009:443).

This view is supported by Schifter (1996a), who contends that a problem posing and problem solving environment as part of a constructivist perspective in learning mathematics, places thinking about mathematical ideas high on the agenda. Indeed thinking about mathematics is valued more highly than memorising algorithms and using them to get correct answers, with a greater onus on individuals to take control and ownership of their learning. Children need to understand …

> that in their mathematics lessons it is *up to them* to offer their thoughts about the questions that are posed. When faced with contradictions to their own conjectures, it is *up to them* to find resolution. (Schifter, 1996b: 79; my emphasis).

Fosnot offers a similar view, emphasising that teachers need to allow learners to raise their own questions, generate their own hypotheses and to test them for viability; her underlying point being that learning is not the result of development but rather learning is development (Fosnot, 2002: 29)

The learning process and the training of thinking in education that Dewey and Whitehead promote, place an emphasis on cognitive development and what it means to understand to ensure knowledge learned can be used and transferred to new applications and novel contexts. These two sub-sections of cognitive development and understanding will be developed before giving further consideration to the many forms of constructivism that are documented in the literature.

2.2.1. Cognitive development

Piaget viewed children as constructing their own learning, seeing the child as a little scientist, exploring the surrounding environment as he or she encountered the world. He considered learning as the development of concepts, and how conceptual knowledge related to experiences as well as linking new knowledge to other concepts. The basic tenet of Piaget's work is that of 'cognitive matching', with learning experiences tailored to the right level for an individual learner. Constructivists view each new experience as adding to existing knowledge, making a growing body of knowledge that Piaget described through a biological metaphor: he observed that to respond to changing conditions, organisms either adjust slightly or undergo a structural change. This is how he perceived learning, using the terms *assimilation* and *accommodation* to reflect the different styles of development that either evolved within existing cognitive structures or forced a structural cognitive change. To make sense of new knowledge, learners could assimilate a fact, skill or technique through a process of adaptation when it was very similar to what they already think and do; the new knowledge does not force a change in thinking or restructuring of existing conceptual frameworks. If, on the other hand the new concept or technique cannot readily be assimilated, then Piaget's accommodation term comes into play with the learner having to alter or displace current thinking to accommodate that new knowledge. Piaget's Theory of Equilibration explains the learning process as a combination of assimilation and accommodation. When doing more assimilating, a child is in a state of equilibrium; when doing more accommodating a child is in a state of disequilibrium. Piaget maintained that organisms are always seeking a state of equilibrium, so this drives the process of accommodation and moves learning forward. The process of accommodation is a more demanding but necessary step in order for learning to be progressed. Learners, who incorrectly try to assimilate knowledge when accommodation is strictly required, are delaying the inevitable and may well encounter deeper difficulties before true progress in learning can be achieved. Piaget's references to assimilation and accommodation reflect his perceived underlying structure to learning and conceptual development that is also supported by others (Byers and Herscovics, 1977, Skemp, 1987). Millar (1989) makes a distinction between concept learning and concept change, with the former understood as a reconstruction of meaning (assimilating as necessary) rather than simply the accretion of new ideas. Concept change on the other hand requires a shift in position, moving from prior understandings to accommodate new experiences. This type of accommodation of new ideas can at times be difficult, with change coming as a result of interaction between existing conceptions and new ideas; the accommodation phase is demanding on the learner who has

to let go of existing beliefs that can be well-established and hard to cut loose to make way for the new conception.

Skemp regards the organisation of knowledge using conceptual structures or schemata as a major feature of intelligent learning. He favours the reference to conceptual structures to emphasise the two qualities in combination, where structure is key to learners' development and integration of concepts and knowledge. Skemp (1987: 119) describes how the structure can be made up of different links or connections that he references as *associative* and *conceptual*. Intelligent learning is progressed when conceptual links (C-links) are made, providing greater scope for generalising and developing a conceptual map. Associative links (A-links) are less efficient, with learners unable to generalise or extend the idea beyond identification of an association. An example of how this would manifest itself can be outlined using a 'number patterns' topic where a sequence is continued or extended. Making A-links would be like continuing the pattern, seeing the connection between subsequent terms but unable to generalise the relationship; C-links would be akin to extending the pattern and being able to offer a general relationship or formula to describe the pattern in full. In terms of learning styles, if A-links dominate then rote learning and memorisation will be witnessed, whereas if C-links are in the majority then conceptual understanding prevails.

The level of cognitive development that can be demonstrated by learners is dependent on the number and strength of connections they have generated. Through such connections, whether structured as a vertical hierarchical network or as a web of concepts and processes, learners can begin to demonstrate extent of learning. Hiebert and Carpenter (1992) view learning as a generative procedure with knowledge constructed by individual students rather than provided by their teacher. They also claim that deep understanding of concepts, ideas and procedures support recall and transfer of knowledge, contending that if cognitive development is built on strong connections and networks of knowledge then fewer individual elements need to be memorised. Similarly, the very nature of construction and re-construction of knowledge-as-a-process places less emphasis on memory-as-storage, and correspondingly makes it easier for learners to recall, use and apply knowledge in new contexts. Hiebert and Carpenter argue these features of connected learning and cognitive development mean one only has to remember 'big ideas' as part of a conceptual network to trigger recall of the component parts, using patterns and connections within personal constructs to prompt recollection of required knowledge and understanding.

2.2.2. Understanding

The question of what understanding means in this context of constructivist learning has been described in various ways by researchers (Byers and Herscovics 1977; Buxton 1978; Skemp 1987), who draw attention to different levels of understanding in an attempt to provide a meaningful framework that educators and learners can use to support cognitive development. Before considering those levels and structures that describe understanding, I want to comment on concepts and processes that can be regarded as components of understanding. Hiebert and Carpenter (1992) discuss conceptual and procedural knowledge where the former embodies rich relationships of connected networks and the latter involves sequences of actions that are replicated in rote fashion with minimal connections for the learner. Students need both procedural proficiency and conceptual knowledge of relationships in order to progress their understanding. Conceptual knowledge and conceptual learning are referenced throughout the literature yet the terms are not always clearly defined. Duckworth (1987 cited in Confrey 1990) identifies qualities that distinguish conceptual understanding from merely completing tasks. She discusses different types of knowledge required for understanding, and defines conceptual knowledge as the giving of names for the way objects are represented as ideas, images, words or formulae. Recall of knowledge is clearly a necessary skill for using and applying knowledge, but it is a skill that is very dependent on how successfully learners structure, organise and understand their knowledge.

Byers and Herscovics (1977) offer three different kinds of understanding that they categorise as: *instrumental* understanding (the ability to apply an appropriate remembered rule, without knowing why the rule works); *relational* understanding (the ability to deduce specific rules or procedures from more general mathematical relationships); and *formal* understanding (the ability to connect mathematical symbolism and notation with relevant mathematical ideas and to combine these ideas into chains of logical reasoning). These three kinds of understanding represent qualitatively different outcomes, stemming as they do from different experiences, with an increasing level of sophistication and mathematical understanding as learners go from a rote expectation to one where formal reasoning, logic and mathematical argument are expected. In terms of communicating understanding, logical reasoning may not equate with logical understanding as there can be flaws in the presented argument, but learners who engage with mathematical proof certainly extend their learning beyond *relational* understanding. This development in level of understanding is akin to a progression from convincing oneself, to convincing others through formal reasoning and explicit articulation of understanding. For example when considering mathematical proof

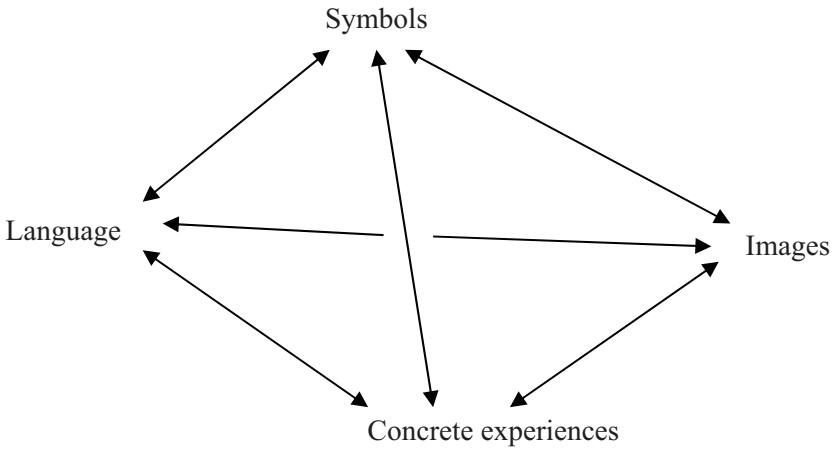
and its place in school curricula, Waring (2000) emphasises the increasing levels of understanding as learners progress through graduated levels of proof, going from personal understandings that convince oneself, to verbal communications to convince a friend, through to a written statement to 'convince a pen-friend'. The latter written statement is considerably more demanding for the learner in that there is no scope for discussion, clarification or embellishment but rather is dependent on inner thought and deep understanding of the concepts in order to convince the reader through a formal written explanation or proof.

Skemp (1987) draws a similar distinction between instrumental and relational understanding of mathematics and supports a comparable structure to that presented by Byers and Herscovics (1977), although Skemp includes a formal level that he calls 'logical'. Most emphasis however, is placed on the differences, distinctions and effects that emanate from the two main categories of instrumental and relational understanding. Four different types of 'knowing' are documented (Yoong, 1987, Mason and Spence, 1999) with distinctions made between 'knowing that' (facts), 'knowing how' (techniques and skills), 'knowing why' (able to justify and reconstruct actions as necessary) and 'knowing to act' (by using and applying knowledge). Skemp's categorisation of instrumental understanding is equated with 'knowing how', and knowing both 'how' and 'why' is analogous with his relational understanding. Skemp argues that learners who pursue instrumental understanding without relational learning and understanding, will have a diminished educational experience and are thereby less likely to think about how knowledge of concepts and procedures can be used in unfamiliar situations; much as argued by Whitehead (1929) through his reference to inert ideas. Buxton (1978) offers a similar breakdown, describing four levels of understanding as: (1) rote; (2) observational; (3) insightful; and (4) formal that emphasises logical or standard form as well as the inclusion of mathematical proof. The first three levels relate to the instrumental/relational categories as discussed above.

These distinctions in the type of understanding that learners develop are particularly important in mathematics and science education where there are clear lines of development, with linear progression within curricular topics and many interconnections between content areas. Learning is enhanced when connections are made within curricular topic developments, across topics within the discipline, and interconnections between content areas including cross-curricular links. Haylock and Cockburn (2003) argue that more secure understanding is achieved where experiences are more strongly connected. Their model of connections draws on the links between different experiences that support the building of cognitive networks and lead to conceptual understanding of a topic. This model builds on

earlier work by Liebeck (1984), who argued abstract thought and mathematical understanding calls for: *experience* with physical objects (E), spoken *language* that describes that experience (L), *pictures* that represent the experience (P), and written *symbols* that generalise the experience (S). The sequence of abstraction tends to follow that order, with 'E' and 'L' preceding 'P' and 'S'; the latter elements more often encountered in textbooks and visual media. When learners are actively engaged in mathematical activity they will typically employ some or all experiences with concrete materials, language, symbols and images. The possible connections between experiences, as shown by arrows in Figure 2-2, might contribute to the understanding of any STEM concept.

Figure 2-2: Significant connections in understanding mathematics (Haylock and Cockburn, 2003)



If learners choose to consider individual aspects of mathematical development separately, this places a significant emphasis on memory and recall, and overlooks the integrated features of mathematics and the benefit of viewing results in a connected fashion. This limited form of understanding is what Skemp (1987), and Byers and Herscovics (1977) refer to as instrumental understanding, where it is still possible to do the mathematics but only by applying remembered rules and having no knowledge of why the rule works. Such an approach can make it quicker to achieve the intended goal, with the learner focused on practising the routine and remembering the concept without having to spend time making sense of the underlying mathematics or its relationship with other activities or concepts. The contexts of learning can often be cited as justification for teaching with instrumental understanding in mind. For example, with a strong emphasis on performance in tests and exams that follow a predicted format. Although instrumental understanding can be quicker to grasp it can also be very limiting for learners.

Internal representations of connected networks and schema are personal and as such can only be shared through communication of thoughts (using discussion, spoken word, writing, pictures, or symbols); in essence, internal representations can only be inferred from communication offered, and a degree of interpretation by the observer. In order to make sense of what someone understands, external representations must be interrogated through appropriate interactions to unpack thoughts, mental connections, schema and networks of knowledge. Internal representations can never be observed. This inability to observe an internal representation of understanding sets apart the behaviourists (Thorndike and Skinner) from constructivist theories and knowledge claims concerning understanding of mental processes.

2.3. Constructivism

Constructivism as a grand theory of learning provides a range of perspectives on how learners actively construct their understanding and knowledge of concepts through experiences and reflection on those experiences. Various shades of constructivism have been documented, each with a set of principles that guide how learners arrive at their personal constructs. Geelan (1997) acknowledges that the field of constructivism can appear very complex to the novice, primarily due to the large number of variants and associated terms of reference that describe perspectives that are very similar in practice. In this section I will first outline the roots of constructivism and conceptual development and then discuss a number of those perspectives, including 'trivial' and 'radical' constructivism, and key features of 'social' and 'cultural' constructivism within educational contexts. In conclusion, I provide an interpretation of constructivism that begins to address the reconciliation of active student-centred learning methods and making sense of mathematics and science disciplines that are essentially bodies of consensually agreed knowledge; a key concern for mathematics and science communities who view the constructivism as used in education to be skirting around the actual learning of an established body of knowledge (Solomon, 2002).

2.3.1. Conceptual development

Confrey and Kazak (2006) identify the notion of misconception as central to constructivist theory. Although 'misconception' is the term often used in the literature there are probably better ways of expressing this concept in any discussion of constructivist theory. For instance calling something a misconception has the connotation of that thought being wrong, whereas this may not be true to the facts if a student's thoughts are as yet not fully developed, but built upon limited experiences to date – their prior learning. Such prior

learning may just be incomplete, with recognised scope for further refinement and development. For this reason, Ausubel (1968) preferred the term 'preconception' to reflect the embryonic and partial state of formation. A distinction is therefore made between 'errors' and 'preconceptions' or 'alternative conceptions', with a greater focus placed on the latter that has a strong bearing on learning and conceptual understanding as discussed in section 2.2.2. Errors are like misconceptions insofar as they result from non-random applications of rules based on certain beliefs, but they differ from misconceptions in that 'errors' tend not to be well-connected to a theoretical or grounded position. This can be witnessed in the work of Sierpinska (1994) who set out four mental operations involved in conceptual development and understanding, and van Hiele (1999) who provides five levels of understanding spatial concepts. Swan (2006: 82) cites the four mental operations depicted by Sierpinska (1994) as:

- *Identification*: we bring the concept to the foreground of attention , name it, and describe it
- *Discrimination*: we can see similarities and differences between this concept and others
- *Generalisation*: we can see general properties of the concept in particular cases of it
- *Synthesis*: we can perceive a unifying principle

Preconceptions can be articulated when we are able to discriminate between that concept and others, but when unable to draw generalisations that fully encapsulate the concept. For example in considering the concept of multiplication, an early conclusion and preconception on the basis of initial experiences with natural numbers might be to state that 'multiplication makes quantities bigger'. This claim distinguishes the concept from other number operations ... until confronted with multiplication by a rational quantity between 0 and 1. Further examples and experiences with rational multipliers will lead to an improved statement, conclusion and synthesis of properties that hold for multiplication of rational numbers. Further refinements and generalisations can evolve when experiences go beyond rational numbers to account for integers, matrices, vectors etc. in order to reach a satisfactory conceptual understanding of multiplication in a variety of contexts.

A similar argument is presented for partial progress within van Hiele's levels of geometric reasoning. The five van Hiele levels are:

1. *Visualisation*: we can name and recognise shapes by their appearance, as total entities, but cannot specifically identify properties of shapes
2. *Analysis*: we begin to identify properties and learn to use appropriate vocabulary that relate to those properties
3. *Informal deduction*: we recognise relationships between and among properties of shapes and can use logical arguments in support of such relationships (e.g. defining a minimal set of properties that uniquely identify a given shape; drawing conclusions on the basis

of known properties of shapes such as: *all* squares are rectangles, but not all rectangles are squares)

4. *Deduction*: we go beyond identifying characteristics to construct formal deductive proofs using axioms and definitions, and show an understanding of the geometric system. For example we can reason about similarity of triangles and can demonstrate the necessary and sufficient conditions for similarity.
5. *Rigour*: we can analyse and compare various deductive systems, such as Euclidean and Turtle Geometry and, know, understand and describe properties of any given shape. The most rigorous level of thought that can comfortably handle abstract mathematics.

(van Hiele, 1999)

These levels are regarded as sequential, so a partial conceptual understanding is demonstrated when learners show signs of reaching say, level 2, at which point they may be considered to have preconceptions about the properties of shape. Partial understanding or preconceptions can be claimed when learners have not as yet reached the final stage of development. Only through further experiences, discussion and *accommodation* of latest practice will a more sophisticated conceptual framework be finalised, to move beyond the preconception status.

Swan (2006) argues that alternative conceptions are a natural and important stage of conceptual development. He further contends that learners and practitioners need to embrace these alternative conceptions as necessary points of transition and that rather than seeking to avoid 'misconceptions' they should confront inconsistencies in a direct manner through carefully structured examples. Brousseau (1997) developed a theory of 'epistemological obstacles' to describe this phenomenon and argued that these alternative conceptions are unavoidable and resistant to change. He makes a case for having epistemological obstacles as a central part of learning, building on preconceptions and supporting learners to become aware of inherent deficiencies in their conceptual frameworks to facilitate understanding and development of improved conceptual frameworks. Brousseau's theory is based on a definition provided by Duroux (1982) who proposed a list of necessary conditions for the term 'epistemological obstacle', namely:

1. An obstacle is a piece of knowledge or a conception, not a difficulty or a lack of knowledge.
2. This piece of knowledge produces responses which are appropriate within a particular, frequently experienced, context.
3. It generates false responses outside this context. A correct, universal response requires a notably different point of view.
4. This piece of knowledge withstands both occasional contradictions and the establishment of a better piece of knowledge. Possession of a better piece of knowledge is not sufficient for the preceding one to disappear. It is therefore essential to identify it and to incorporate the reason for its rejection into the new piece of knowledge.

5. After its inaccuracy has been recognised, it continues to crop up in an untimely, persistent way. (Brousseau, 1997: 99-100)

Given there is a resistance to change preconceptions, as evidenced in the literature (Ausubel, 1968; Biemans and Simons, 1995; Driver, 1983), Brousseau's fourth point above is fundamental to successful conceptual development. Ausubel's oft cited comment that preconceptions are 'amazingly tenacious and resistant to extinction' stems from the fact that personal constructs have to be changed to accommodate subsequent experiences. One or two classroom activities are not going to change strongly held beliefs. Driver (1983) argues that learners need time, in groups and with their teacher, to think and talk through the implications and possible explanations of what they have observed and experienced. Open discussion and talk will be required to take adequate account of prior knowledge and how that knowledge is organised and constructed. Only then can learners build on what they already know to accommodate new knowledge, and for new beliefs to be accepted.

Cognitive conflict can be a stimulus for learning, and can determine the organisation and reorganisation of what is being learned. These conflicts arise from personal experiences, as highlighted by Dewey, who noted that reflection and subsequent actions were instigated by the sources of cognitive and conceptual conflict: 'Reflection arises because of the appearance of incompatible factors within an empirical situation. Then opposed responses are provoked which cannot be taken simultaneously in overt action' (Dewey, 1916: 326). Whitehead (1967:93) also made the point that making mistakes should not be conceived as a disaster; arguing that students 'must be free to think rightly and wrongly' and that error should not be hidden or explained away. Fosnot and Perry (2002) and Swan (2006) reiterate that point, acknowledging the importance of error in conceptual development, seeing error as a unique opportunity and an additional source for gaining new knowledge through talk, discussion and argumentation.

2.3.2. Trivial and radical constructivism

Two basic principles of constructivism, cited by von Glasersfeld (1989), are:

1) knowledge is not passively received but actively built up by the cognizing subject;
2) the function of cognition is adaptive and serves the organization of the experiential world, not the discovery of ontological reality.

The simplest version of constructivism, coined 'trivial constructivism' by von Glasersfeld, is when only the first of these two principles is observed. Geelan (1997) categorised this perspective as 'personal constructivism', where the focus is on individuals as

they construct knowledge for themselves, drawing on Piaget's processes of assimilation and accommodation for concept development. Acceptance of the second principle is more controversial in that such a stance challenges the epistemology of the discipline as a subject domain (mathematics, science or any indeed any other discipline). This somewhat revolutionary stance represents an extreme view of constructivism that von Glasersfeld termed 'radical constructivism'. Kilpatrick (1987) considers radical constructivism from a developmental perspective, claiming that constructivism could describe the 'evolution of child thought' without challenging epistemology of mathematics (or science). Kilpatrick was certainly opposed to von Glasersfeld's radical constructivism if it meant rejection of an external, knowable reality and objective truth that encapsulated the Platonist view of mathematics. Kilpatrick summarises his views on 'radical constructivism' as:

> ... an epistemology that makes all knowing active and all knowledge subjective. Following modern physical sciences in its rejection of the possibility of coming to know ultimate reality, it treats the cognizing subject as the organizer of his or her own experience and the constructor of his or her own reality. It views coming to know as a process in which, rather than taking in information, the cognizing subject through trial and error constructs a viable model of the world.
> Kilpatrick (1987: 5)

The choice or wording, through using 'viable model of the world' for learners constructing their models of reality, stresses the ontological difference between this model and any claim for reconstruction of objective truth.

Jaworski (1993) presents a comparable interpretation of radical constructivism as applied to the learning of mathematics. She contends that *if* there is some independent, pre-existing body of mathematical knowledge, we cannot know it other than through our own experiences; we can only know what we have witnessed through experience and constructed for ourselves, with modifications and updates in light of further experiences. This interpretation does not dismiss the possibility there is some pre-existing body of mathematical knowledge, indeed it assumes such a Platonic stance by including knowledge created by mathematicians, but Jaworski claims 'we can never know it in any absolute sense'. She also stresses the epistemological nature of constructivism makes no ontological claim as a theory of learning.

2.3.3. Social and cultural constructivism

Airasian and Walsh (1997) and Geelan (1997) both draw distinction between the personal and social perspectives of constructivism. Personal, or Piagetian constructivism as discussed above, outlines a developmentally organised cognitive construction that ignores the socially and historically situated nature of knowing in favour of new knowledge that is

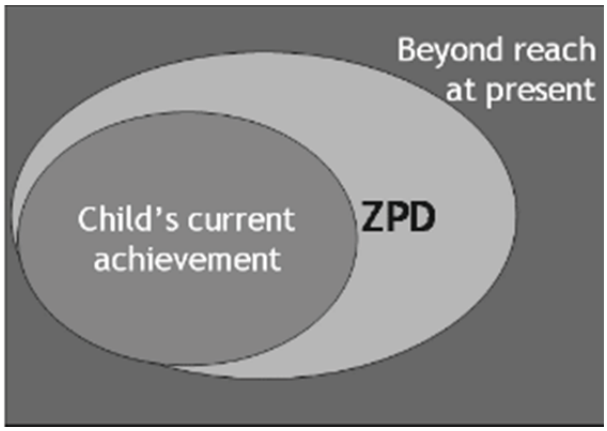
constructed internally or metacognitively. In contrast, the social component is central to knowledge development in social or sociocultural constructivism that is also referenced as Vygotskian constructivism.

Social constructivism takes forward ideas similar to those promoted by Piaget in his theory of cognitive development but with a stronger emphasis placed on the social aspect of learning. In *Thought and Language* (Vygotsky and Kozulin, 1986) the arguments in support of social constructivism are set out; originally published in 1934 and also known under a variant title of *Thinking and Talking* that emphasises the central tenets of the theory. Vygotsky views language and thought as interdependent for learners, with young children moving from thinking out loud towards internalising their thoughts. He argues that this transition from egocentric speech to a point where language is used to describe and articulate ideas, demonstrates learning and cognitive development. The social aspect of sharing and developing understanding through talk and discussion with 'more knowledgeable others' (peers, adults, teachers) is regarded as central to learning for Vygotsky. The concept of cognitive development being enhanced or accelerated through social and cultural experiences is strongly attributed to Vygotsky and his support of language as a means to learning with and through others. Vygotsky (1978) discusses how learning can be extended when individuals work with more knowledgeable others, as opposed to pursuing tasks solely on their own. For instance one might start out unable to do a particular task or understand a body of knowledge but accelerated progress can be made with the assistance of another, be that a teacher or just a more knowledgeable peer. This process of supporting learning is often presented in the form of a Venn diagram (using set theory as illustrated by Atherton (2011) in Figure 2-3) where unaided learning, children's current achievement, is bounded. The scope of potential development with the aid of 'more knowledgeable others' is defined as the zone of potential (or next) development, commonly translated as 'zone of proximal development'; both abbreviating to ZPD. Vygotsky (1978) defines the ZPD as the distance between the 'actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers' (1978: 86). Cognitive psychologists have perhaps focused too much on trying to understand internal mental representations, structures and schema at the expense of wider socio-cultural influences that may be of equal importance.

In Geelan's summary of the many forms of constructivism, he highlights the 'social' dimension of social constructivism. For Geelan (1987: 18), knowledge firmly resides in communities and societies rather than within cognising individuals or awaiting discovery

within the natural world. This perspective on knowledge is therefore reliant on dialogue and discussion within communities.

Figure 2-3: Vygotsky's Zone of Potential Development (ZPD) (Atherton, 2011)



Knowing how others conceive of mathematics and science is a challenging enterprise (Steffe, 1995) and one that can only be progressed through social and cultural interaction to ascertain what beliefs and concepts are held, and how they might be extended or modified in light of new experiences and contexts. These social and cultural interactions, that share meanings and share knowledge, relate to the way we approach learning mathematics and science. Fosnot (1996) argues that knowledge does not exist outside a person's mind and because constructivism entails individual constructs, one can never say whether or not two people have produced the same construct. The best that can be claimed is that the constructs are compatible, and seem to function in the same way as each other. Through discussion and exposure to further experiences, those models can become closer to each other to arrive at some consensual point.

Vosniadou (1996) outlines 'situativity theory' as the natural development from the work of Piaget and social constructivism of Vygotsky, focusing as it does on the interaction between social agents and the physical environments in which they live and learn with individual cognition being socially and culturally mediated. My main emphasis is placed on cultural factors and experiences that are more likely to have a direct bearing on learning, cognitive development and students' achievement. Socio-cultural influences are important later in life when knowledge is to be recalled and *used* in genuinely authentic situations and contexts. The argument therefore is to secure learning with understanding that can be transferred readily to applications in everyday life and applied to new contexts and applications beyond school. School learning is itself embedded within a culture, a school culture that can be very different from everyday culture. When there is a mismatch between

the ways STEM is taught in school and the way STEM is used in real life situations, knowledge will not be transferred to out-of-school contexts. Vosniadou (1996) concludes that if wider cultural influences on learning are not adequately addressed, then school learning will have little to do with the culturally relevant and will consequently run the risk of being abstract and irrelevant to learners; akin to 'inert knowledge' referenced by Whitehead. The social and cultural dimension provides a framework for learning, with like-minded communities working together and constructing their knowledge in relation to that held by the community. According to Wenger (2006) these 'communities of practice' share a common interest or goal (*domain*), work in a supportive and collaborative manner (*community*), and actively pursue similar experiences, problems and challenges (*practice*) en route to learning and developing an understanding of their domain. School and classroom settings provide these three structures for learners and educators to form such a community. One of the distinguishing features of 'communities of practice' is the way that learning is enhanced within that community of adults and children, through their collaborative engagement on common activities and purposes – their common culture (Rogoff, 1994). Correa-Chavez and Rogoff (2005) argue that cognitive and social development is supported through such participation in learning, where children and others learn through observation and being part of the process. The role of expert people in communities of practice is perceived to be different from that articulated by Vygotsky. Vygotsky's 'more knowledgeable others' are expected to provide the lead and to explicitly extend the learning for individuals by taking them into their ZPD; in communities of practice it is the 'experts' collaborative participation that is important. Correa-Chavez and Rogoff see the expert contributors providing a desirable culture for learning rather than necessarily providing explicit pointers in reaching a solution or conclusion for the shared task. This style of learning is referenced as *intent participation* to reflect the highly focused observation and contribution made by learners who take responsibility for their own learning; the motivation coming from being invited to participate and contribute to a valued communal activity. In contrast, Rogoff *et al.*(2003) refers to *assembly-line instruction* where responsibility for learning resides mainly with the experts who manage the learners' experiences by giving them bite-size tasks that lead to the whole; learners do not have the big picture revealed until further into a sequence of lessons. Collaborative activity between experts and learners is not a feature of this alternative model of learning.

Bereiter (1994) discusses *knowledge building* as distinct from learning activity and recognises that all knowledge cannot necessarily be personally constructed. There is more to learning STEM than mastery of knowledge, including aspects that are reliant on a social and

cultural construction of knowledge. Inculcating learners into the processes of doing science and mathematics is such an important feature that it cannot solely rely on being implicitly covered. For example Harlen (2006) asserts there is scope for explicit promotion of the scientific method and how scientists and mathematicians operate in real life situations. Driver *et al.* (1994) present similar arguments for socially constructed knowledge, where 'learners need to be given access not only to physical experiences but also to the concepts and models of conventional science'. There is no claim from the social or cultural constructivists that all aspects of STEM education require individual construction, but they would argue individuals must make sense of the ways in which knowledge claims are socially constructed, validated and communicated (Driver, 1989; Driver *et al.*, 1994; Bereiter, 1994).

The distinguishing features between the various shades of constructivism, from trivial to radical via social and cultural constructivism, focus on the processes of construction and how interaction with others can influence and guide personal constructs and models of reality in a world view where we can only know what we have constructed. Promotion of the scientific method and a scientific culture will include the opportunity to actively experience question-raising, hypothesising and predicting; observing, planning and conducting investigations; interpreting evidence and communicating findings. Central to such activity is talk; these are not solitary actions but rather all benefit from collaborative approaches.

2.4. Talk and Argumentation

Barnes *et al.* (1978) provide a seminal text on speech as communication and as reflection where ideas are transformed into words. Much student-student and teacher-student talk provides opportunities to clarify thinking and to learn from one another. Atwood *et al.* (2010) explore the merits of different types of talk as categorised by Mercer (1999, 2006, and 2008). The three categories of talk, identified by Mercer as exploratory, cumulative and disputational, involve varying levels of reasoning and claim to have a correspondingly variable impact on students' learning. Exploratory talk offers the greatest potential for learning given the emphasis it places on learners to justify their position and to publicly articulate their understanding. This public airing of personal understandings includes the questioning of assumptions, outlining reasons for claims and addressing challenges from others in a positive and constructive fashion. Exploratory talk is ultimately a cooperative interaction that leads to collective development either through consensus or by arriving at a point where others can take things forward by refining or developing ideas

under consideration. Cumulative talk is also consensual, but differs from exploratory talk in that there is little or no scope for challenge or justification of claims. Participants agree on a shared understanding of knowledge without seeking external validation or scrutiny of claims, relying instead on being accountable only to one another. Cumulative talk may appear to represent cooperative interaction, but Atwood *et al.* (2010) assert its lack of critical evaluation and public accountability make it a less effective form of talk than the exploratory form. Although cumulative talk can be positive and supportive for participants, the lack of explanation or justification make it insufficiently searching or challenging to take forward learning as effectively as exploratory talk. The third category, disputational talk, tends to be defensive and oppositional. In this form of talk participants will hold entrenched views and primarily rely on opinion rather than any substantiated support for claims. In general, disputational talk fails to generate explanations or evidence of listening to alternatives. A consequent result is that there is no collective understanding or genuine collaboration in disputational talk, a feature that limits the extent of knowledge development and learning. Mercer (2008) concludes that exploratory talk offers the greatest benefits to learners, citing evidence from his study of over 700 children where successful problem solving was noted alongside targeted exploratory talk. Those children exposed to the intervention of exploratory talk 'made significantly greater gains on math and science tests than did those in control classrooms' (Mercer, 1999: 107). The research claim on exploratory talk is that agreed common understanding created by a group is greater than that with which individual participants started (Wells and Aruz, 2006; Atwood *et al.*, 2010). The airing and sharing of diverse, and at times conflicting perspectives has a positive influence on learning where there are demands made to explain, clarify and justify claims through appropriate dialogue.

Alexander (2006) also extends student-talk into asking questions and to the central role of 'argumentation' that involves explaining and evaluating thinking and understanding. Argumentation or argumentative discourse is the social process of pursuing a particular line of thought or 'argument'. Kuhn and Udell (2003) offer a concise distinction between these terms:

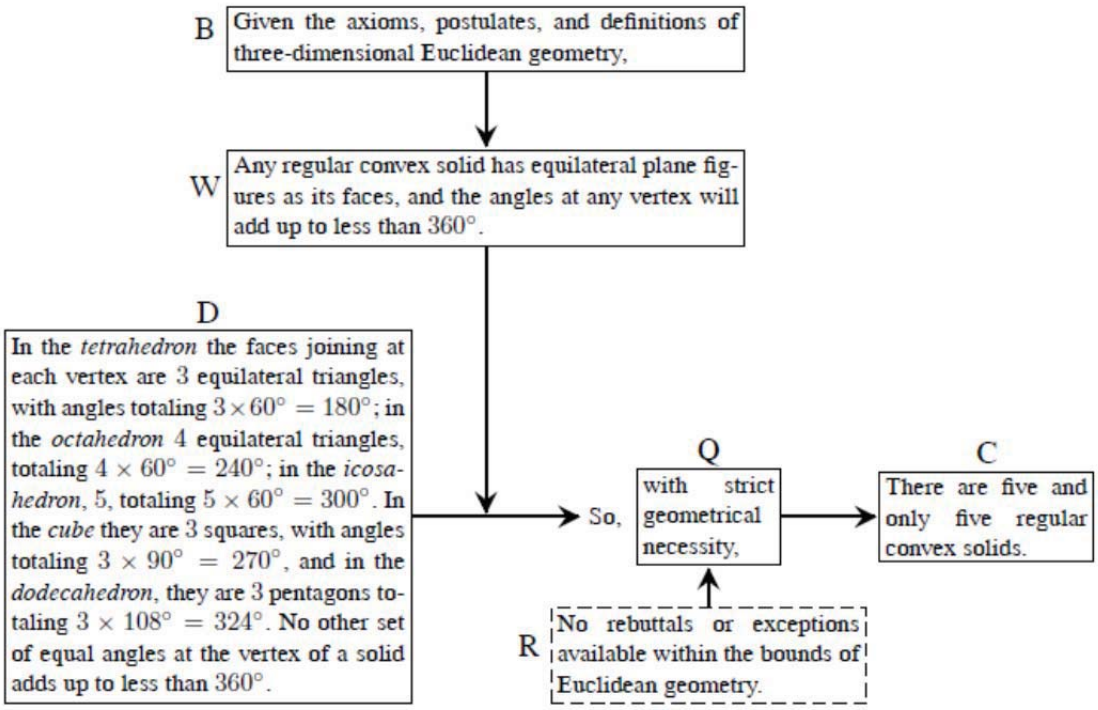
> The terms argument and argumentation reflect the two senses in which the term argument is used, as both product and process. An individual constructs an argument to support a claim. The dialogic process in which two or more people engage in debate of opposing claims can be referred to as argumentation or argumentive discourse to distinguish it from argument as product.
>
> Kuhn and Udell (2003: 1245)

Jiménez-Aleixandre and Erduran (2007) suggest that argumentation can be

considered at two levels. The first is at an individual level where an argument is 'an inner chain of reasoning' within an individual's mind as they develop and articulate that point of view. The second is the social meaning identified above where differences are resolved through debate with an externalisation of internal thinking and reasoning. Jiménez-Aleixandre and Erduran (2007) make a strong case for argumentation being introduced in the science classroom, presenting five interrelated potential contributions to learning that result from its introduction. The first contribution to learning comes through making cognitive processes public. Argumentation calls for collaborative approaches, drawing from situativity theory and communities of practice, with participants expected to back up their claims with evidence and to evaluate alternative options or explanations in a public manner; potentially beneficial to learning. The second contribution comes through developing communicative competencies and critical thinking skills. As discussed earlier, the importance of language and communicative skills is clearly acknowledged within a socio-cultural theory of learning. The relevance of language in knowledge construction is captured in science learning through the notion of 'meaning making' (Mortimer and Scott, 2003), through promotion of scientific language, the appropriation of scientific discourse, and engaging with scientific method. The commitment to evidence that argumentation demands, runs parallel to the disposition of critical thinkers who pursue evidence for their beliefs; again, potentially beneficial to learning. A third contribution to learning stems from the promotion of scientific literacy, pursued through talking and writing science. This focus on the spoken word and written language in science learning is partly in response to 'recipe-style' experiments and laboratory work, and an over-emphasis on procedural components of scientific knowledge that is memorised in a rote format rather than being learned or conceptually understood. Through argumentation, students develop scientific literacy by learning to talk and write the languages of science, and become appreciative of the practices of science, its culture and epistemology. Students' knowledge of scientific method, where members of the community 'propose, justify, evaluate and legitimise knowledge claims within a disciplinary framework' (Jiménez-Aleixandre and Erduran, 2003: 9) can support students' learning and more generally their enculturation into the practices of science and what it means to be a scientist. There is clearly a commitment to evidence when promoting argumentation but depending on personal philosophical standpoint and understanding on the nature of science or mathematics, there may be variance over what counts as evidence in this world of evidence-based arguments. Introducing argumentation into science classroom practice certainly raises these questions and issues for learners to grapple with; an exercise that Jiménez-Aleixandre and Erduran suggest as potentially beneficial to learning.

The importance of argument in education beyond science is set out by Andrews (2009) who asserts academic discourse within schools and academia would be nothing without argument. Andrews contends the act of social exchange is part of the benefit, with argumentation providing a means to: clarify personal position and understanding; recognise alternative or opposing views; and to defend espoused position. He goes on to claim the process offers catharsis for participants and the potential to discover ideas that can persuade self and others of validity in stated claim and supporting evidence. The reference to 'self' relates to Vygotsky's reflection and argumentation, whereby inner argumentation is preceded by thinking aloud and the public defence of a stated claim: 'all that is internal in higher mental functions was at one time external' (Vygotsky, 1991:33). As an individual articulates a point of view, they can be said to be developing an argument. At a naïve level, such talk can be construed as thinking aloud and 'talk for your own understanding', before progressing to 'talk for someone else's understanding', an activity that calls for explanation and successive shortening of process. A final destination might be to record that shorthand notation to have a written explanation or proof, fit for a 'pen friend' as suggested by Waring (2000). Andrews (2009) cites the social psychologist Vygotsky and philosopher Habermas in support of argument being at the heart of academic discourse, and draws upon Toulman's Argument Pattern (TAP) as a supportive framework for educational settings. In essence, an argument comprises a claim and evidence in support of that claim. This simple model that is expanded upon by TAP, to go from Data (D) to a Claim (C), taking account of supporting evidence in the form of a Warrant (W), with Backing (B) in support of that warrant. The claim may also be qualified in some way through stating a Qualifier (Q) with any consequent limitations acknowledged by noting an exception or rebuttal (R). To illustrate the uses of such an argument, Figure 2-4: Toulmin's analysis of Theaetetus' proof that the platonic solids are exactly five in number (Aberdein, 2005: 291) is reproduced below in accordance with the logic originally presented by Euclid. Toulmin's warrant, W, combines the definition of a regular convex polyhedron with the essential prior result on which the proof depends, aspects that are propositions in Euclid's Elements (Book XI); most of Euclid's proof is contained within Toulmin's data , D, that sets out the finite cases to be considered.

Figure 2-4: **Toulmin's analysis of Theaetetus' proof that the platonic solids are exactly five in number (Aberdein, 2005: 291)**



Andrews (2009) concludes that argumentation plays an important role in learning, noting the opportunity it affords to learn from challenges, errors and misconceptions; and from presenting evidence that eliminates alternate hypothesis on the grounds of reason and logic.

2.5. Practical activities

In the early part of the 20th Century the main purpose of practical work was viewed as a way to develop conceptual understanding. Following publication of the Thomson Report (1918), practical work was justifiable only where it offered support to the learning process, supporting conceptual understanding and rejecting the previously held view that scientific discovery was paramount in school science, and where practical and investigative methods dominated proceedings. In subsequent decades there was a shift to focusing on physical practical skills that were transferrable to technological industries (1940s). This rise in individual investigative activity placed time constraints on what could be studied, but in the 1950s it was a narrowing of the curricular syllabus that gave way to those time constraints rather than any reduction of practical work or indeed a transfer of practical activities from individuals to teachers, where demonstrations of practical work could have been more time-

efficient. In the 1960s, a rise in popularity of 'pupil-as-scientist' saw a need for learners to 'do science' in order to 'understand science', but from the1970s onwards there have been mounting doubts over whether 'doing' does in fact lead to enhanced understanding (Driver, 1983; Hodson, 1996). The life mantra: 'I hear and I forget. I see and I remember. I do and I understand.' (Confucius) is called into question by Driver as she reflects on the potentially negative impact of practical work in school science learning. Driver (1983:9) favours 'I do and I am even more confused', as a more likely outcome from much of what is witnessed under the guise of practical work in school science.

Practical work of itself is not going to enhance learning, but given the right conditions, teachers' support and appropriate pedagogic skills in the practical elements of science, practical work can contribute to learning in different ways. Hodson (1996) suggest five purposes for practical work in science education:

1. To enhance the learning of scientific knowledge
2. To teach laboratory skills
3. To develop 'scientific attitudes' such as open-mindedness, objectivity, willingness to suspend judgement
4. To give insight into scientific method and develop expertise in using it
5. To motivate pupils by stimulating interest and enjoyment.

Of those purposes, only the first leans towards conceptual understanding, while the next three relate to skills and scientific processes (knowledge of scientific method) and the last concerns the affective domain and motivational impact on learners who participate in practical work. These broad purposes of science education are widely supported in the literature (Shulman and Tamir, 1973; Anderson, 1976; Hofstein and Lunetta, 1982; Abrahams, 2011). Each categorization recognises three distinct strands that impact on students' engagement with practical work in science – conceptual understanding; scientific method and associated skills; and affective value of practical work.

There is often a taken-for-granted acceptance that practical work is 'good', with the tacit belief that learners' conceptual understanding is enhanced through such activity. Practitioners do practical work because it is 'expected' within science education, but in so doing perhaps overlook the breadth of the key purposes cited above (Hodson, 1996). However, empirical data from studies of practical science do not endorse the claim that students' engagement with practical work enhances their conceptual understanding of the subject. Major reviews of the literature and research have all concluded that the presence of practical work offers no significant benefit in terms of students' scientific conceptual understanding (Hofstein and Lunetta, 1982, Clarkson and Wright, 1992). Clarkson and Wright note that this is an uncomfortable conclusion for something that is commonly considered to be invaluable, although it may be down to the form of data capture that is used

within the reported studies:

> despite the frequent claims that one purpose of practical work is to provide an effective means of developing conceptual understanding, research findings suggest, at least when outcomes are measured using pen and paper tests (and you might like to consider whether this approach is justified), that there is no significant advantage (or disadvantage) to its use.
>
> (Abrahams and Reiss, 2012:1037).

Conceptual understanding comes as a result of connections learners make as they construct their knowledge (Hiebert and Carpenter, 1992). The processes that learners have experience of, including scientific experiments, investigations and practical tasks, will determine the extent of understanding that is achieved. It is more a matter of *how* those experiences are developed, with a strong emphasis placed on learning rather than merely 'doing' practical work. Abrahams and Reiss (2012) criticise the use of recipe-style practical activities that fail to make the required links with big conceptual ideas. For teachers who do not have a science specialism, having students 'doing' pre-defined practical tasks may be useful, in that the recipe-style practical is well-defined and unlikely to generate extended discussion and explanation; thereby making fewer demands of the teacher. However, learners who are restricted to 'doing' recipe-style practical tasks miss out on a depth of learning that comes from co-constructing, discussing and evaluating experiments in relation to theory.

White and Gunstone (1992) advocate a three-step model that shifts students from 'doing' to 'learning' where they are expected to 'Predict – Observe – Explain' (POE) when undertaking practical work. The initial predict-phase calls for a justification of the stated prediction, so learners immediately have to engage with the practical activity and its concepts as they think about options and plausible outcomes. This is followed by a description of what happens in the observe-phase and then any conflict between prediction and observation must be reconciled in the third and final phase – 'the explanations students proffer in this step reveal(s) much about their understanding' (White and Gunstone, 1992: 46). The extent that learners are called upon to explain their work is central to practical experience. Explaining practical work begins to make bridges between the observations and conceptual ideas.

The more opportunities learners have to move away from the recipe-style practical, the stronger their understanding will become. Developing sound experiential knowledge of the scientific method and associated skills, ideally an un-sanitised version of scientific method, will be what takes learners' understanding forward. Abrahams (2011) suggest that learners need to experience situations where experiments and practical work does not always work out as planned or anticipated according to the published 'recipe'. Science is often less

certain, with the relationship between experiment and theory at times in conflict; a feature that learners need to witness in order to develop a full understanding of the scientific method. Laboratory skills are also developed in practical experiences where the learner 'becomes a scientist' and appreciates the breadth of skills associated with 'being a scientist'. Content specific skills such as using instruments, for example accurately reading and interpreting scales and measures using burettes, or performing techniques like titration, can only be developed through hands-on experience. Content independent skills like problem solving and various aspects of scientific method on the other hand could be developed within and beyond the science laboratory and are certainly not confined to practical tasks.

The third strand of practical work that impacts on students' engagement with practical activities lies within the affective domain. Students generally enjoy practical work and certainly have a *relative* preference over non-practical, theoretical or written activities. Abrahams and Reiss (2012) report on students' views on practical work, noting that almost all respondents liked practical work. However, when probed further it was apparent the reasons students provided for liking practical work were either relative to other learning activities or as an absolute preference, with the latter offering reasons that were grounded in practical tasks. Only 32% of students expressed an absolute preference for practical work, where they liked it on merit rather than just being better than non-practical activities e.g. absolute preferences included descriptors like: fun; do things; see what happens; find things out; exciting; gain experience. When analysed by age/stage it is clear that as students get older, the proportion reporting absolute preferences falls away, from Yr 7: 54% absolute; 46% relative to Yr 8: 26% absolute; 74% relative – a big shift in favour of relative preferences for the older students (Yr 8). This trend continues as students extend their studies and progress to Yr 10. Of those who do offer absolute preferences, fewer than 6% give reasons that relate to learning, understanding and recollecting such as: will remember it better; learn more; helps you understand better.

The reasons given for liking practical work provide information on students' motivation or interest in that aspect of their learning. The motivational value of practical work appears to offer short-term engagement rather than any longer lasting intrinsic appreciation of and enthusiasm for science. This motivational role of practical work making learning interesting is important for learners and teachers to capitalise upon. According to Hidi and Harackiewicz (2000), a distinction needs to be made between personal and situational interest. Situational interest is local to the practical activity or context, an interaction that is instigated by the particular situation to motivate the learner. If the interest is situational, then that motivational input needs to be continually provided for each new

situation, experiment, practical task or context. In contrast, personal interest in a topic or subject has a longer term effect on how one approaches learning, with students' intrinsic motivation for a subject rising above the particular. Whatever practical tasks offer by way of interest and motivation, it is important to consider the hook that this pedagogy provides. Hidi and Harackiewicz (2000) draw parallels with Dewey's terms of 'catch' and 'hold', whereby practical work can capture students' interest at one level, but another higher order benefit comes when practical work can hold students' interest over time and context. They cite groupwork, puzzles and ICT as a way of stimulating and 'catching' students' interest, but these organisational approaches and resources fail to maintain or 'hold' students' interest over time. More engaging and deeper organisational and pedagogical strategies will be more fruitful in 'holding' students' interest. For example giving autonomy to learners with opportunities to *design and plan* investigations and to engage in meaningful tasks, have proven to be empowering variables that maintain students' interest and are more enduring over time and context.

A student-survey undertaken by Cerini, Murray and Reiss (2003) asked 14-19 year old students about their expectations of science education. The study covered a broad cross-section of the cohort with nearly half of the respondents already signed up to and studying science, the remaining students either not currently or having no expectation of studying science. The students were presented with 11 ways of learning and asked to select three possibilities that they find *most useful and effective* in helping understand school science. A second question sought to elicit their three preferences on methods of teaching and learning that they found the *most enjoyable* part of school science.

The findings are reported in Table 2.5-1, with the ways of learning ordered by proportion of support for 'most useful and effective'. Respondents were clear that what they most enjoyed wasn't always what they reported as most useful and effective. Having a discussion or debate is noted as being the most useful and effective form of learning and it is also regarded as enjoyable by nearly two-thirds of respondents. Doing a science experiment was in the top three for being useful and effective for learning, and as reported in other studies discussed earlier, doing experiments is regarded as enjoyable by over 70% of respondents. The form of data collection, where respondents select from these 11 ways of learning, does not offer further insight to how science experiments or investigations are pursued. For instance we do not know if there is scope for students to design and plan these practical tasks; if they have moved from 'doing' to 'learning' when undertaking practical work; or if they can show an understanding of the ideas they were expected to engage with. The highlighted data in Table 2.5-1 draw attention to the ways of learning that are (a) valued

in terms of effectiveness but not enjoyed – taking notes from the teacher; and (b) enjoyed but not particularly valued by the learners as being useful and effective – making a science presentation in class. On the basis of my earlier discussion on the merits of explaining and making presentations, one might expect these two features to be reversed in terms of effectiveness and benefit for learning. It is worth noting that the students enjoy making those presentations, but just tend not to value presentation as a way of learning relative to others on the shortlist provided.

Table 2.5-1: Students views on 'effective' and 'enjoyable' activities (Cerini *et al.*, 2003: 10)

Ways of learning	Useful and effective (%)	Enjoyable (%)
Having a discussion / debate in class	48	64
Taking notes from the teacher	45	15
Doing a science experiment in class	38	71
Doing a science investigation	32	50
Going on a science trip or excursion	30	85
Looking at videos	27	75
Taking my own notes from books etc.	24	13
Copying notes from the board	23	17
Reading the textbooks	17	18
Making a science presentation in class	17	43
Researching science on the internet	8	4

n=1,450

When this cohort of 14-19 year old students were asked to reflect on their primary school experiences, and to suggest how primary school science might have been different, over half of the respondents recommended an increase in practical activities. The data shown in Table 2.5-2 suggest that students believe they must see what is going on in science – either through personal practice or observation (demonstration or video); an increase in use of IT is not strongly supported.

Table 2.5-2: Students' reflective view on Primary School Science

At primary school science should:	
Be more practical	56%
Have more theory	10%
Be more visual (videos etc.)	28%
Be more IT based	6%

n=1,464

Practical work has also been justified in terms of addressing behaviour management and ways of working with lower ability students. There are spin-off benefits of reduced

class size for practical classes (maximum of 20 in Scotland) and staff can find themselves using the practical as a carrot for improved behaviour, knowing the students like to do practical work. Abrahams (2000) cites several cases of science teachers using such an argument as part of their student-behaviour management strategy, where the carrot of practical work provides an additional option for the teacher concerned. As far as lower ability students are concerned, he cites teachers who view practical tasks as something 'to do' with those learners. One respondent summarised her practice as:

> They're so weak they can't even do the calculations, you know, so they can't even plot the graph, you know, it's [practical work] just something to do with them

Another respondent reiterated similar sentiments as he commented:

> It [practical work] gives them something to do, especially the ones that get bored with too much writing (Abrahams, 2000: 43)

White (1979) highlights the importance of practical work in providing an effective anchor for subsequent recall and application of underlying concepts. This goes beyond the interest-catching mode described above by incorporating memorable practical events into students' learning episodes. These may be memorable due to being visually spectacular, by presenting a novel approach, or by making strong connections to existing knowledge through appropriate contexts. White argues that having memorable practical episodes in conjunction with sound pedagogical practice can provide the necessary links to students' knowledge and skills, further supporting their constructivist learning:

> in addition to providing suitably unusual experiments which will establish episodes in students' minds, teachers will have to consider how to encourage them to link the episode with appropriate intellectual skills and verbal knowledge. (White, 1979: 386).

White concludes that those memorable events can be experienced in three types of experiment or practical activity. First there will be the spectacular and unusual experiment that can be factored into learning a few times in any one year, providing strong links for recall of the most important topics or themes of study in a programme. Second there are more routine experiences that provide learners with meaningful contexts and offer potential links between school studies and real life applications of STEM education. The third type of experiment is one that requires genuine action from the learner, where problem solving exercises call for students to integrate and extend their knowledge of STEM subjects. The power of problem solving as a means of enhancing learning in this way and the strong influence of constructivism in STEM education is further supported in the review compiled

by SCORE partners (2008) who call for authentic and meaningful problems to catch and hold students' interest and to provide opportunities for 'learners [to] construct knowledge by solving genuine, meaningful problems' (Lunetta *et al.*, 2007).

The authors' key findings include reference to *effective pedagogy* and *subject-specific professional development*:

> *Effective pedagogy* is at the heart of improving the quality of practical work in science. When well-planned and effectively implemented, practical work stimulates and engages students' learning at varying levels of inquiry challenging them both mentally and physically in ways that are not possible through other science education experiences. (Key Finding 5, SCORE 2008:13)

> *Subject-specific professional development*, or rather the lack of it, has been highlighted in other reports. More specifically the questionnaire responses indicated that, although 21% of teachers engaged in CPD specifically related to practical work in the last year, over 40% indicated they could not remember 'ever' receiving CPD on practical work. Opportunities for training and professional development for teachers and for technicians, to support practical work, need to be improved and teachers and technicians engaged with these. (Key Finding 11, SCORE 2008:17)

These findings summarise the specific issues at stake in STEM education where learners need leadership and direction from well-qualified and confident practitioners. Teachers need to critically reflect on key elements of practical work and effective pedagogy in STEM, paying due attention to both the effective and affective value of practical work. Students need to have increased opportunities to interact with ideas, pushing for links between practice and theory and generating evidence of conceptual understanding. This will probably result in less time merely 'doing' practical work.

2.6. Scottish policy context

Scottish education policy was in a period of transition when the data for $TIMSS_{2007}$ was collected with the Scottish Government's flagship *Curriculum for Excellence (CfE)* being implemented across schools and local authorities. CfE was launched with a publication of the curriculum review group's report and ministerial response (Scottish Executive, 2004c & 2004b). This initial report captured the big ideas and underlying principles that CfE sought to promote with a raft of subsequent publications providing finer details on curricular content and teaching approaches. In order to make appropriate links to learning and teaching as outlined in this chapter, the curriculum and assessment policy papers and their antecedents related to the aims of the curriculum in Scotland and approaches to learning and teaching are outlined.

The curricular aims of CfE are to ensure that all children and young people in Scotland develop the knowledge, skills and attributes to flourish in life-long learning, and to

make a positive contribution to society through work and citizenship. Those aims and purposes are presented as an overarching model that addresses the oft-quoted *four capacities* that enable children and young people to be 'successful learners, confident individuals, responsible citizens and effective contributors' (Scottish Executive, 2004b). A deeper analysis of those capacities highlights attributes and capabilities contained therein, reflecting much of what has been gleaned from recent research on learning theories as discussed earlier in this chapter as well as building on the best of earlier Scottish curricular policy developments. Design principles of particular note in CfE are the references to breadth, depth, challenge and application in the general documentation with further exemplification in subject-specific guidance on pedagogical approaches as outlined in the *Principles and Practice* series (Education Scotland, 2010a & 2010b). Although that particular series of policy guidance for mathematics and science was not published until after the survey data collection, many of the constructivist teaching approaches recommended were already present in the existing policy framework of national guidelines for Curriculum and Assessment in Scotland for 5-14 (Henderson and Cunningham 2011). The 5-14 national guidelines for mathematics (SOED, 1991) and science within environmental studies (SOED, 2000), provided the policy framework prior to CfE (Scottish Executive, 2004b). Many of the key messages within those guidelines were reinforced in the intervening years between publication and 2004 through various evaluations of effective learning and teaching, standards and quality reports, and inspectorate reviews including *Improving Achievement* series (SOED, 1993; SOEID, 1996a, 1996b, 1997a, 1997b; Scottish Executive, 1999; HMIE, 2001a, 2001b, 2005a, 2005b, 2006; Learning and Teaching Scotland, 2004 . McNab (1999) notes that the SOEID reports, *Achievement for All* (1996a), *Achieving Success in S1/2* (1997a), and *Improving Mathematics Education 5–14* (1997b), indicate that the Inspectorate in Scotland view 'levels of mathematical attainment as less than satisfactory and in need of serious improvement'; hence the plethora of reviews and reports in that short time period as the policy makers sought to pursue, consolidate and further develop what they viewed as effective practice within the policy framework. Around the same time as $TIMSS_{2007}$ was administered and reported, further national guidance on CfE emerged (Education Scotland, 2010) as well as updated progress reports and evaluation of learning and teaching in Scottish schools. These later publications included a new version of *Improving Scottish Education* (HMIE, 2009), *Science: A portrait of current practice* (HMIE, 2008), *Learning Together: Mathematics* (HMIE, 2010) and *Excellence in Mathematics* (Scottish Government, 2011b). Those publications supported policy implementation by highlighting current strengths of the system and identifying approaches to learning and teaching that were deemed in need of

further development.

Assessment policy within CfE was developed from the work of Black and Wiliam (1998) and the natural evolution from the *Assessment is for Learning* (AifL) programme started in 2002. When the Scottish Executive published *Ambitious Excellent Schools* (Scottish Executive, 2004a) it set out a target of having every school in Scotland committed to the principles of AifL by 2007, to ensure that assessment supports learning. The principles that make an AifL school centre on 'learning together' with assessment as an integral part of learning and teaching. The three strands of assessment – *for*, *as* and *of* – learning, as outlined in Chapter 1, are underpinned by four principles, where students:

1. understand clearly what they are trying to learn, and what is expected of them;
2. are given feedback about the quality of their work, and what they can do to make it better;
3. are given advice about how to go about making improvements;
4. are fully involved in deciding what needs to be done next, and who can give them help if they need it.

Hattie and Timperley (2007) identify feedback as one of the most powerful influences on learning and achievement although Hattie's synthesis of meta-analyses of influences on student achievement (Hattie, 2009) highlighted the variable impact that different types of feedback can have on students' achievement. Highest effects were associated with receiving feedback about a task and how to do it more effectively; lower effect sizes were related to extrinsic motivational feedback related to praise, rewards, and punishment. Hattie and Timperley (2007) make a distinction between feedback about the task (FT), about the processing of the task (FP), about self-regulation (FR), and about the self as a person (FS), arguing that FR and FP are the most effective forms of feedback in that they focus on deep processing and mastery of tasks. In the *AifL toolkit*, FR relates to assessment *as* learning; FP is aligned with assessment *for* learning; and FT and FS are broadly associated with assessment *of* learning where the feedback component is crucial to plan for improvement – providing qualitative detail that goes beyond a simple mark or score.

The transformational change-in-practice sought by the CfE policy can only be achieved if everyone involved, including students, practitioners throughout the system, and policy makers, subscribe to the four principles set out above and pull together in the interests of learners. Initial analyses of assessment practices in the Highland Council of Scotland (Hayward, Spencer and Simpson, 2005) support earlier findings by Hallam *et al.* (2004) and Condie *et al.* (2005) in that teachers, heads of establishments, and local authority coordinators were all strongly convinced that the AifL project was highly effective. Hayward *et al.* (2005) identified that changes in teachers' assessment practices were

perceived to have been happening, but questions remained over why practices were changing and what implications there might be for AifL as it was scaled up. As with the curricular policy documents referenced above, the timing of data collection for $TIMSS_{2007}$ preceded publications on assessment within CfE. The overarching principles of assessment were set out in a strategic vision paper (Scottish Government, 2009a) and more detailed policy recommendations on assessment followed in the *Building the Curriculum* (BtC) series of papers; the most pertinent being *BtC_5: Framework for Assessment* (Scottish Government, 2011a). The principles of assessment and latest policy guidance are based on experiences within the AifL project that appears to have taken hold in the system as a whole. The emphasis that is placed on assessment to support learning was consistent with the values and beliefs of teachers; the integrity of this practice is a necessary component for effective change in a system, where those that implement the policy believe that what they are doing matters (Fullan, 2003; Hayward *et al.*, 2005).

In discussions of curriculum and assessment policy implementation, and approaches to learning and teaching in mathematics and science education, there is an emphasis on the need for a skilful mix of approaches, including:

- **Reform-based practice and discussion**

  Reform practices reflect the social constructivist development in STEM education that has at its core an emphasis on problem solving and enquiry. The concomitant shift in practice places conceptual understanding as a primary goal with discussion and talk to the fore. HMIE (2005a) reported teachers to be in pursuit of focused questions to check on students' understanding rather than simply eliciting correct answers; they further noted students having to explain their thinking in line with policy aspirations. Through effective questioning and discussion in mathematics, teachers can also use misconceptions as opportunities to build upon and deepen students' understanding of mathematical concepts. Improving mathematics education 5-14 (SOEID, 1997b) notes that all teachers of mathematics should have their students spend more time in explaining and asking questions to demonstrate depth of understanding, with less class time spent on individual practice. Opportunities for discussion and explanation of thinking are also a feature of science education in Scotland, with the need for students to explain their observations during science practical experiments and investigations; much as advocated by White and Gunstone (1992) where students are expected to 'Predict – Observe – Explain' (POE) when undertaking practical work. Discussion and debate, expressing informed opinions and making decisions on a range of issues, are presented as central tenets of developing students as scientifically literate citizens:

'Discussion was best when the teacher asked open-ended questions, designed to make students think about scientific concepts and to allow them to acquire and practise the vocabulary and language of science'(HMIE, 2005b: 20). The current analyses of $TIMSS_{2007}$ data will provide evidence of impact from those practices, identifying patterns of association with achievement scores that will support he claims or refute the practice-as-implemented to be non-significant.

- **Active learning and practical activities**

  Active learning is promoted as learning that engages and challenges children's thinking using real-life and imaginary situations. This provides opportunities to observe, explore, investigate, experiment, play, discuss and reflect with a focus on cognitive activity as well as doing practical tasks. All Scottish teachers are encouraged to adopt more interactive and participative methods that engage pupils actively in learning (SOEID, 1997b). Teaching through enquiry and investigation is promoted to develop problem-solving capabilities and critical thinking skills. Learners should be encouraged to think deeply about mathematical ideas and concepts, to deepen their own understanding and to apply their skills and knowledge in a range of contexts as they work on familiar, unfamiliar and non-routine problems. Raising expectations on ownership of tasks and decision making equally applies to science education where policy initiatives dictate investigative skills are developed and students undertake practical experiments. The three-step model in the guidelines of 'prepare – carry out – review and report' demands an increasing level of autonomy as students strengthen their role in preparing or planning experiments and investigations; the review and report element should also increase in scale as students develop their investigative skills. The 5-14 National Guidelines for science, make a case for students doing practical experiments, where it states: 'First-hand investigations are central to the way in which young children learn science, providing opportunities to plan fair tests, make observations, hypothesise, predict, collect evidence, research, survey and discuss. Through such means, opportunities arise to infer, deduce, calculate, draw conclusions from evidence, make judgements and debate important issues.' (SOED, 2000). Given the data on planning, doing and watching practical experiments and investigations, an evaluation of those claims and policy initiatives can be made, reporting on differences across stages as students progress their science education.

- **Learning environment**

  Scottish policy dictates there should be opportunities for learning collaboratively and independently, noting that learning can be most effective when there are opportunities

to think and talk together to share understanding and clarify misconceptions as learners move towards relational and conceptual understanding. At all stages, it is claimed an emphasis on collaborative learning will encourage children to reason logically and creatively through discussion of mathematical ideas and concepts. HMIE (2005a) note missed opportunities for students to learn collaboratively, and consequently missing out on further developing their understanding of mathematical ideas. In the secondary stages, HMIE (2006) call for more collaborative approaches that will engage students more actively in thinking about their learning as well as requiring them to develop their creative thinking skills. Analyses of $TIMSS_{2007}$ will provide evidence of practice and association with achievement to comment on the effectiveness of collaborative study across disciplines and stages.

- **Contexts**

  Using relevant contexts and applying knowledge and understanding to new situations is highlighted in the curricular guidance, encouraging links across the curriculum to illustrate how mathematical concepts are applied in a range of contexts including those in science and social studies. HMIE (2005a) noted staff in most primary and secondary schools had not as yet fully considered the impact that different learning and teaching contexts could have for developing students' learning skills and confidence for applying mathematical skills; however this particular feature did not get classified as a 'main area' for improvement. There are noteworthy overlaps with active learning and students' confidence and security on subject knowledge when it comes to students' exposure to contexts. A broad range of familiar and unfamiliar contexts can provide good opportunities for students to be suitably challenged in their learning, to demonstrate depth of knowledge, and to apply their understandings with confidence – cross referencing to the underlying principles of CfE where depth, challenge and application are highlighted. Survey data can provide a national picture of opportunity, impact and strength of association of contexts with achievement in mathematics and science education to support or refute HMIE claims based on selected inspection reports.

- **Appropriate and effective use of ICT**

  HMIE (2005a) report some schools to be making good use of information and communication technology to enliven learning and teaching, but this is not consistently reported across the system. The most common weakness reported in primary schools concerned students' skills in using ICT, with specific reference made to their lack of understanding, knowledge and skill in using databases and spreadsheets. The *improving*

series publications reported secondary teachers beginning to make very effective use of ICT to enhance learning and increase the pace of learning through their use of interactive boards. However, students at all stages are reported to make too little use of ICT in secondary science (HMIE, 2005 and 2008), particularly where there should be good opportunities in collecting, analysing and presenting scientific data. $TIMSS_{2007}$ data will provide a national profile of ICT as used in P5 and S2, with analyses identifying whether ICT is associated with achievement as claimed in the research literature and policy documentation.

- **Assessment and feedback**

  The policy framework sets out a central purpose of supporting learning within assessment practices, going on to suggest this is best achieved through a combination of formative and summative assessment. Assessment should probe the ability to apply the learning in challenging tasks and in unfamiliar situations. Learners will demonstrate success by building on previous learning and being able to make links in learning by looking back as well as forward and through using feedback to good effect. The timely nature and quality of feedback is claimed to facilitate students' progress as they experience assessment in a range of ways. Formal and informal assessments including assessment of on-going classwork, reviewing homework, working on a quiz or test, and completing examination items for reporting purposes can all contribute to formative feedback. Analyses of survey data will provide evidence of impact, with the following elements from the *AifL Toolkit* highlighting practices that should support such developments:

  • ensure that any feedback is positive and encouraging but that it always points towards a specific action for improvement.
  • students take responsibility for, and are active in, their own learning.
  • every student can make progress from where they are, based on assessment and feedback of their last piece of work/activity.

The Scottish policy context has clearly changed in the way that curricular content is communicated through the statements of *outcomes* (I can …) and *experiences* (I have …) but the constructivist practices and reform teaching that is promoted within CfE documentation has been very much an evolutionary development from earlier learning and teaching approaches that were captured in the Cockcroft Report (1982) and 5-14 National Guidelines (SOED, 1991 & 2000). Given the reform practices have been promoted over a number of years and that the key messages have been well-documented in published reviews and

evaluations of practice, it is reasonable to expect evidence of such practice being reported in the $TIMSS_{2007}$ data. A question therefore remains over whether those reform practices and students' experiences reported through the survey questionnaires are associated with achievement, and if so, how the strength of association can be measured and used to guide and inform future practice.

2.7. STEM Education – claims

Learners' experiences provide the focus for analyses of the survey data. The operational variables in this particular study will be aligned with theories of learning and teaching that can be categorised as 'traditional' or 'reform', where the latter seeks to reflect recent and current initiatives within Scottish education policy and schooling. Analyses of achievements and experiences documented in $TIMSS_{2007}$ will provide empirical evidence of practices and their association with achievement. The findings will report on relative effectiveness of those learning experiences and their association with a 'reform' agenda that claims to enhance and progress education, and improve school effectiveness.

Respondents provide evidence of those experiences through specific responses to student-, teacher- or school-level questionnaires that are directly or indirectly associated with key aspects of learning and teaching. This includes specific experiences that relate to: practical and active approaches to learning; situations that offer a greater or lesser extent of student autonomy in the learning process; and organisational aspects of the learning environment. Some experiences are reported by both student and teacher, in which case the student-variable is used to reflect the learners' perspective on that aspect of learning and teaching; where no student-variable was recorded, the teacher-variable is used as evidence on that particular experience. Most of the categorical variables are reported on the basis of *'Every or almost every lesson – About half the lessons – Some lessons – Never'*, with the categorical variables for science G4 reported on: *'At least once a week – Once or twice a month – A few times a year – Never'*. To facilitate interpretation and to make communication of findings more meaningful, a reverse coding is deployed so that the predicted effect of high achievement is evidenced by high levels of engagement with any specified experience; a minor methodological transformation is applied to the responses.

Drawing on the reported learners' experiences, the literature and policy claims can be investigated to ascertain levels of association with achievement and to determine whether any patterns of association are empirically consistent within the hierarchical structure of the data. Variances at student-, teacher-, and school-level are reported through the multilevel models to provide commentary on consistency of practice and association with achievement on a national basis. The evidence presented in the literature review informs the direction of association expected for each of the operational variables, with the reform agenda of STEM education making claims for enhanced learning opportunities and implicit claims for raising achievement outcomes. The anticipated direction of association for each variable in the empirical analyses is as set out in Table 2.7-1 to Table 2.7-4, with cross-reference to the related learning theory or practice that has informed this stance.

Table 2.7-1: TIMSS G4 experiences in MATHS lessons

	Student Variable (S) Teacher Variable (T)	*Description*	*Anticipated association with achievement*
	Reform-based practice and discussion		
1	AS4MHEXP	Explain my answers	Positive – with links to literature on social constructivism, relational understanding, exploratory talk, argumentation, and collaborative learning; these are all advocated as desirable features within reform-based practice.
2	AS4MHMWP	Memorize how to work problems Memorize formulas and procedures	Generally Negative – where behaviourism, instrumental understanding, inert ideas, and associative links are less well supported when learning reduces to memorisation of processes. Clearly a level of memorisation is desirable where 'facts' and building blocks of understanding are concerned but excessive memorisation connotes with rote learning.
	Active learning and practical activities		
3	AS4MHMCL	Measure things in the classroom and around the school	Generally Positive – social and cultural constructivism, collaborative learning, affective value, but this type of experience should be complementing other approaches to study and should not dominate a student's classroom experience.
4	AS4MHTCG	Make tables, charts or graphs	Generally Positive– social and cultural constructivism, collaborative learning, constructionism and application of learning by producing an output that will communicate findings form problem or investigation. Association may reverse if either of the above is in excess, reflecting a possible lack of challenge, low expectations and correspondingly lower levels of achievement.
	Learning environment		
5	AS4MHWSG	Work with other students in small groups	Positive – links to literature on collaborative learning, social constructivism, talk, discussion and argumentation. Reform-based practice would imply this to be a desirable an essential feature of learning
6	AS4MHWPO	Work problems on my own	Generally Positive – given the opportunities to consolidate learning and understanding through completion of individual tasks. Potentially negative if this was a sole approach to learning.
	Contexts		
7	AT4MASDL	Relate what they are learning in mathematics to their daily life	Positive – links social and cultural constructivism, relational understanding, and heuristic methods in problem solving. Applying knowledge to new situations in a range of contexts should provide depth of learning and strengthen understandings.

Table 2.7-2: TIMSS G8 experiences in MATHS lessons

	Student Variable (S) Teacher Variable (T)	*Description*	*Anticipated association with achievement*
	Reform-based practice and discussion		
1	BS4MHEXP	Explain my answers	Positive – links with social constructivism, relational understanding, exploratory talk, argumentation, collaborative learning; as in Table 2.7-1 (Variable 1) these are all advocated as desirable features within reform-based practice.
2	BS4MHFRR	Memorize formulas and procedures	Generally Negative – where behaviourism, instrumental understanding, inert ideas, and associative links are less well supported in the literature. Response here infers memorisation of processes and procedures, reducing the experience to one of rote learning that has been widely criticised in literature review for its limiting impact on students' ability to apply their understandings.
	Active learning and practical activities		
3	BS4MHSCP	Decide on our own procedures for solving complex problems	Positive – links to literature on social and cultural constructivism, collaborative learning, heuristic methods of problem solving and especially autonomy and ownership of learning that has been commended and explicitly promoted in policy developments.
4	BS4MHGCT	Interpret data in tables, charts or graphs	Positive– with links to social and cultural constructivism, collaborative learning, and argumentation. Interpretive and critical thinking skills are developed through opportunities to make sense of data (applications of learning), enhancing depth of learning and strengthening understandings.
	Learning environment		
5	BS4MHWSG	Work together in small groups	Positive – collaborative learning, social constructivism, talk, discussion and argumentation as justified in Table 2.7-1 (Variable 5)
6	BS4MHWPO	Work problems on my own	Generally Positive – as noted for G4 mathematics in Table 2.7-1 (Variable 6). Same qualification of being potentially negative if this was a sole approach to learning.
7	BS4MHLSP	Listen to the teacher give a lecture-style presentation	Generally Negative - where high levels of lecture-style presentation at the expense of other reform practices reflect a traditional 'chalk-and-talk' approach that can be passive for learners and counter to reform-based practice. However, students' responses on this experience may present an alternative picture if their interpretation of 'lecture-style' is in line with 'whole-class-teaching' that may be very interactive; in this case a positive association with achievement would be anticipated with the reform practice of interactive learning opportunities.

Table 2.7-2: TIMSS G8 experiences in MATHS lesson (Continued)

Student Variable (S) Teacher Variable (T)		*Description*	*Anticipated association with achievement*
	Contexts		
8	BS4MHMDL	Relate what they are learning in mathematics to their daily life	Positive – social and cultural constructivism, relational understanding, heuristic methods of problem solving as justified in Table 2.7-1 (Variable 7)
	Assessment and feedback		
9	BS4MHROH	Review our homework	Positive – with links to literature on feedback, discussion, relational understanding, conceptual links in understanding, and use of exploratory talk.
10	BS4MHHQT	Quiz or test	Positive – if pursued in the spirit of AifL where the emphasis is on formative processes, this links to literature on affective value, feedback, discussion, relational understanding, conceptual links, and exploratory talk. On the other hand the experience of 'tests' may be counter-productive in terms of raising levels of students' confidence and conceptual understanding and will not necessarily be positively associated with achievement. Much as lecture-style presentation was noted as having a potentially negative association, so too could the deployment of 'quiz or test' as it will be dependent on the teachers' approach and implementation – in spirit and in practice.

Table 2.7-3: TIMSS G4 experiences in SCIENCE lessons

Student Variable (S) Teacher Variable (T)		*Description*	*Anticipated association with achievement*
	Reform-based practice and discussion		
1	AS4SWESS	Write or give an explanation for something I am studying in science	Positive – links with social constructivism, constructionism, relational understanding, exploratory talk, argumentation, and collaborative learning; as in Table 2.7-1 (Variable 1) these are all advocated as desirable features within reform-based practice with the additional support for 'explaining' in the context of practical experiments and investigations in science.
2	AS4SMESF	Memorize science facts	Generally positive – given the reference to 'science facts'. A level of memorisation is necessary across all levels of achievement where 'facts' are concerned, but excessive memorisation connotes with rote learning. Where memorisation is extended to processes and procedures, one might reasonably expect a negative association with achievement as cited in Table 2.7-1 (Variable 2).

Table 2.7-3: TIMSS G4 experiences in SCIENCE lessons (Continued)

	Student Variable (S) Teacher Variable (T)	***Description***	***Anticipated association with achievement***
	Active learning and practical activities		
3	AS4SWATE	Watch the teacher do a science experiment	Generally Negative – recipe style, inert ideas, with practical being used as a management strategy and learners potentially passive in the experience. In the absence of students doing practical experiment themselves, watching science experiments is anticipated to provide a positive experience. Similarly where watching provides a memorable event and conceptual understanding, with teacher modelling scientific method, and stimulating intrinsic motivation through a *catch & hold* strategy, one might expect a positive association with achievement.
4	AS4SHPEX	Design or plan a science experiment or investigation	Positive– social and cultural constructivism, conceptual understanding, affective value, constructionism, collaborative learning, heuristic methods of problem solving, engaging in scientific method, argumentation, and most importantly autonomy and ownership of the task that should of itself provide intrinsic motivation (catch & hold)
5	AS4SDESI	Do a science experiment or investigation	Generally Positive – linking to theories on collaborative learning, social constructivism, constructionism, scientific method, talk, discussion and argumentation, as well as research on practical experiments. However, given mixed reviews on the role of practical experiments in the science classroom this variable may offer a negative association, or reduce level of positive association, where practice includes adoption of a recipe-style approach or one that is unconnected to wider science education (inert ideas)
	Learning environment		
6	AS4SHWGX	Work with other students in a small group on a science experiment or investigation	Positive – on the grounds of supporting collaborative learning, social constructivism, talk, discussion (argumentation), scientific method, catch & hold (intrinsic motivation), much as justified in Table 2.7-1 and Table 2.7-2 (Variable 5).
7	AS4SWSPO	Work on science problems on my own	Generally Negative – while this practice may link with Cockcroft's recommendations for 'consolidation and practice of fundamental skills and routines', it misses out on collaborative work that is strongly promoted in wider literature on science education. Same qualification as noted in Table 2.7-1 (Variable 6) with potentially negative association if this was a sole or primary approach to learning.

Table 2.7-3: TIMSS G4 experiences in SCIENCE lessons (Continued)

Student Variable (S) Teacher Variable (T)	*Description*	*Anticipated association with achievement*
Contexts		
7 AT4SCSDL	Relate what they are learning in science to their daily life	Positive – social and cultural constructivism, relational understanding, heuristic methods in problem solving and applying knowledge to a range of contexts.

Table 2.7-4: TIMSS G8 experiences in SCIENCE lessons

Student Variable (S) Teacher Variable (T)	*Description*	*Anticipated association with achievement*
Reform-based practice and discussion		
1 BS4SHEOS	Give explanations about what we are studying	Positive – with links to social constructivism, constructionism, relational understanding, exploratory talk, argumentation, collaborative learning as noted in Table 2.7-3 (Variable 1)
2 BS4SHFAP	Memorize science facts and principles	Generally Positive – as cited in Table 2.7-3 (Variable 2) given the emphasis on facts and principles over processes and that memorisation will be a feature of learning across all levels of ability/achievement. Excessive memorisation may connotes with rote learning, so where memorisation is extended to processes and procedures, one might expect a negative association with achievement as commented upon earlier.
Active learning and practical activities		
3 BS4SHDEI	Watch the teacher demonstrate an experiment or investigation	Generally Negative –with learners potentially passive in the experience. As with G4, in the absence of students conducting practical experiments, watching experiments is anticipated to provide a positive experience. Similarly where watching provides opportunities for social and cultural constructivism, conceptual understanding, or through a memorable observed event, one might expect a positive association with achievement
4 BS4SHPEI	Design or plan a science experiment or investigation	Positive– as reported for same variable in Table 2.7-3, with particular credence given to autonomy and ownership, under the auspices of active learning.
5 BS4SHCEI	Conduct an experiment or investigation	Positive – collaborative learning, social constructivism, constructionism, scientific method, talk, discussion and argumentation. Similar qualifications concerning the style of experiment as noted in Table 2.7-3 (Variable 5).

Table 2.7-4: TIMSS G8 experiences in SCIENCE lessons (Continued)

Student Variable (S) Teacher Variable (T)		*Description*	*Anticipated association with achievement*
	Learning environment		
6	BS4SHWGO	Work in a small group on an experiment or investigation	Positive – collaborative learning, social constructivism, talk, discussion etc. as noted for G4 science in Table 2.7-3 (Variable 6)
7	BS4SHWPO	Work science problems on my own	Generally Negative – as noted in Table 2.7-3 where working on own is a primary approach to learning. Scope for some positive outcomes through 'consolidation and practice' as cited in Cockcroft's recommendations.
8	BS4SHLSP	Listen to the teacher give a lecture-style presentation	Generally Negative – much as noted in Table 2.7-2: TIMSS G8 experiences in MATHS lessons with high levels of lecture-style at the expense of other reform-based practices. As in mathematics, the students' interpretation of 'lecture-style' may equate with whole-class-teaching of an interactive nature, in which this case a positive association with achievement would be anticipated in line with reform practice of interactive learning opportunities.
	Contexts		
9	BS4SHMDL	Relate what they are learning in science to their daily life	Positive – social and cultural constructivism, relational understanding etc. for the same reasons as noted in Table 2.7-2
10	BS4SHROH	Review our homework	Positive – links to literature on feedback, discussion, relational understanding, conceptual links, exploratory talk, and argumentation. Arguments presented for AifL practices and the importance of feedback on work tackled, noting the 'type of feedback' will have a bearing on the final outcomes.
11	BS4SHHQT	Quiz or test	Positive – linking to literature on affective value, feedback, discussion, relational understanding, conceptual links, exploratory talk, and the principles of AifL as outlined in Table 2.7-2. Similar qualification on 'tests' is applicable in science, where they may inhibit levels of students' confidence and will not necessarily be positively associated with achievement.

Models of student's achievement that incorporate those variables will provide empirical evidence of their effect, and will either support or refute the claims stated in literature and policy documentation. Similarities and differences across disciplines and stages will also be highlighted, providing evidence of association between practice and achievement to inform practitioners and policy makers as next steps in curricular review and

development are progressed. Findings will be *indicative* of policy effectiveness, highlighting where more focused studies might be targeted to go beyond the student-perspective and to probe policy initiatives through more direct routes, recognising the limitations of $TIMSS_{2007}$ data that fails to necessarily ask the most pertinent of questions.

3. Methodological issues related to survey design and analysis

A number of methodological issues and methods of data analysis are influenced by the sampling design and data structure of the surveys. Working with secondary data offers distinct opportunities and benefits to analysts who can take advantage of the large-scale data collection that has been rigorously undertaken and compiled into user-friendly data sets. However, the same data presents challenges to analysts in that *their* questions have not necessarily been asked and only indirect measures of problems may be available for analysis. An additional issue, beyond the fact that all desired information may not have been collected, concerns missing data. Data can be missing in surveys through non-response, but also as a result of the survey design implemented by primary researchers. In this section I will first outline the broad issues surrounding missing data before discussing the practical implications of missingness within the design and data collection phases of $TIMSS_{2007}$. Finally I will summarise the theoretical approaches used to analyse the surveys, providing details of a new multilevel plausible value method of analysis that is subsequently compared with a range of existing methods of analyses, each of which will be evaluated on empirical grounds in the latter part of my dissertation.

3.1. Missing data

In its simplest sense, data can be missing as a result of respondents failing to complete what is asked of them. This could be on a major or minor scale. First, a major loss of data comes as a result of a sample unit failing to respond or to complete any of the items included in the survey. The units of interest in $TIMSS_{2007}$ are students, teachers and schools, each of whom feature as sample units in the survey with questionnaires administered at each level. Accordingly, this type of missing data is referred to as *unit non-response*. Second, and arguably a less major situation is where the unit may respond to the questionnaire incompletely. This is referenced as *item non-response*, where some questions, but not all, are answered for a particular unit in the selected sample.

A more complex situation of missing data comes as a result of survey design. In $TIMSS_{2007}$, the data collection is restricted to limit the time and cognitive demands made on individual students. Researchers are not interested in reporting on the particular students (units) in the sample but in the population of students for which the observed students are deemed to be representative. From the researcher's perspective, any student (unit) in the population could have been used to generate the data, and could therefore take the place of

each unit in the observed sample; students are exchangeable. In drawing conclusions on population statistics it is residuals that are of interest, and residuals are exchangeable. A consequence of this design decision is that there will only be imputed data or *plausible values* as cognitive measures or achievement scores for individual students. This situation presents a variant on missing data, with an incomplete data set through design rather than non-response. What is important for subsequent analyses is that the concept and nature of missing data has to be understood and suitably addressed by analysts if survey data are to be evaluated and interpreted in an appropriate manner.

Sinharay *et al.* (2001) identify three ways that missing data arise in studies. These are categorised as 'missing completely at random' (MCAR), 'missing at random' (MAR), and 'missing not at random' (MNAR). Missing data would be categorised as MCAR if the probability of a missing response is independent of all the measured and unmeasured characteristics of the individuals in the study. Responses are categorised as MAR if missing responses do not depend on the missing values; missingness may, however, depend on other characteristics of the individual. In those two cases where data are MCAR or MAR, satisfactory techniques are available for analysts to make good estimates of model parameters. The most problematic type of missingness, in terms of statistical analysis, are those data deemed to be MNAR. Statistical assumptions that rely on the random nature of responses do not hold when it is known that missingness is related to the value sought. For instance, if missing cognitive values were not observed because of a lack of confidence or low self-esteem from a group of respondents, then it can be claimed the missing data are related to the value that would have been observed and the data are 'missing not at random'.

Given an incomplete data set as a result of non-response there are broadly three ways to approach the statistical analysis. First, the analysis could be restricted to complete cases only. This 'complete case' analysis would only use those cases that contained complete information on all of the variables included in the study. Complete case analysis results in a reduced data set that makes the exercise inefficient, but for the statistical analysis to provide valid results it also relies on the assumption that missing cases are MCAR. If that assumption of MCAR is violated the complete case analysis, or listwise deletion method, can lead to biased results. A second approach is to use any available data; deleting cases if and only if the variable of interest is missing. This method of 'pairwise deletion' will not include a particular variable when it has a missing value, but it can still use the case when analysing other variables with non-missing values. Pairwise deletion usually has fewer deletions than 'complete case' analysis, dropping only those cases that are missing values on the particular variables being studied. It will provide valid results only if the MCAR assumption is

satisfied. As above, violation of that assumption will run the risk of analyses being inaccurate. Pairwise deletion, that bases its results on more data than complete case analysis, would appear to offer a more accurate analysis but in some ways this belief is misplaced because different estimates can be based on different cases, making it difficult to have an overall interpretation of analyses that are dependent on correlations within the data. A third approach is to consider substitute data when confronted with an incomplete data set.

3.1.1. Imputation

This practice of imputation, or 'filling in' missing data with plausible values, is widely recognised as an attractive way of handling incomplete data (Little and Rubin, 1987; Rubin, 1987; Schafer, 1999 & 2003; von Davier *et al.*, 2009; Mislevy *et al.*, 1992a). At its most basic, the strategy involves a single imputation for each missing value to generate complete data. These imputed values are inserted to allow analyses to be carried out using the usual statistical techniques for complete data. A range of imputation strategies can be used to identify credible values for the missing data, each with its limitations but overall facility to generate complete data. Typical examples of imputation strategies include:

- *mean substitution* where all the missing values for a variable are replaced by the mean value in the observed data for that variable
  (valid estimate of mean; underestimates variation. Intuitively, the single imputation must underestimate the variance since all missing values are given the mean value with no further variance included.)
- *hot deck imputation* where each missing value is replaced by an observed value from a randomly chosen case that is similar to the case with the missing value
  (preserves marginal distributions; difficult to implement where there is more than one missing variable)
- *regression substitution* where all the missing values of a data set are replaced by the predicted value of that variable from a regression analysis based on the complete cases
  (valid estimate of mean parameters; underestimates variance parameters because the method does not assume a residual error around the regression line)
- *stochastic regression imputation* where all the missing values of a data set are replaced by the predicted value of that variable from a regression analysis based on the complete cases plus a random residual term
  (a better procedure and an improvement on regression substitution)

The first two imputation strategies above do not rely on distributional properties of the data but the latter two above, rely on aspects of the distribution such as the mean and variance parameters that define a suitable model of the observed data.

Sinharay *et al.* (2001) propose further ways of handling missing data and introducing

substitute values through imputation strategies that use either maximum likelihood estimation (MLE) or Bayesian estimation as a means to generate complete data distributions. A key difference between these two strategies and those mentioned above is that the MLE and Bayesian methods estimate the parameters of interest without any requirement to create a complete data set in the first place. They assume data are MAR and use a formal probability model to make inferences on units with missing data. Inference is based on the observed data likelihood that links the observed data and the parameters of interest. It is worth noting at this point that these methods are equally applicable to the second type of missing data, where incomplete data is through design rather than non-response as witnessed in $TIMSS_{2007}$.

3.1.2. Maximum likelihood estimation

For maximum likelihood estimation, the task is to find the global maximum value of the parameter vector for the likelihood function based on observed data. The likelihood function $L(\theta|y)$, represents the likelihood of parameter θ given observed data y ; essentially the reverse of the probability density function that specifies the probability of observing data y given the parameter θ:

$$f(y|\theta) = f(y = (y_1, y_2, \dots, y_n)|\theta) = f_1(y_1|\theta)f_2(y_2|\theta) \dots f_n(y_n|\theta) = \prod_{i=1}^{n} f_i(y_i|\theta)$$

The maximum likelihood estimation (MLE) is a method to find the probability density function (PDF) that makes the observed data most likely. To simplify the computational demands that are associated with that product of probabilities, the ML estimate is obtained by maximising the log-likelihood function, $\ln L(\theta|y)$. Assuming the log-likelihood function is differentiable, the maximum values for θ will be determined from solutions to the likelihood equation:

$$\frac{\partial \ln L(\theta|y)}{\partial \theta_i} = 0$$

where $\theta_i = \theta_{i,MLE}$ for all $i = 1, \dots, k$. In practice, however, it is not always feasible to obtain an analytical solution for the MLE estimate, especially when the model involves many parameters and the PDF is complex. The MLE is therefore more normally found through computer calculations that pursue an approximate numerical solution using iterative procedures.

One currently popular procedure to find solutions for ML estimates when dealing with incomplete data is the Expectation-Maximisation (EM) algorithm. This algorithm, attributed to Dempster, Laird and Rubin (DLR, 1977), is deemed to be efficient in terms of

computer storage and ease of programming, and reliable in identifying global maximums. DLR (1977) presented the EM algorithm as a general technique for finding MLE from incomplete data. The two step algorithm, involves an initial estimate ('E' for expectation) that allows incomplete data to be 'filled in' on the basis of the identified parameters, which in turn permits a complete-data solution to identify a new estimate ('M' for maximisation step). Missing data are then re-estimated using the new parameters to generate a second complete-data solution, repeating in an iterative fashion until the parameter values converge to the point where there is no real change.

When data are missing, we can partition Y as $Y = (Y_{obs}, Y_{mis})$ where Y_{obs} includes the observed components of Y, and Y_{mis} includes the missing components of Y. EM finds the ML estimate of θ from observed data only, the likelihood function for θ given observed data: $\ln L(\theta|Y_{obs})$. The algorithm then iteratively maximises the log-likelihood function, $\ln L(\theta|Y)$, using complete-data methods. If $\ln L(\theta|Y_{obs})$ is bounded, the sequence of $\ln L\left(\theta^{(t)}\middle|Y_{obs}\right)$ values converge to a stationary value of $\ln L(\theta|Y_{obs})$, where the superscript (t) denotes values obtained at the t^{th} iteration. This value is used as input for the next iteration. In summary, each iteration consists of two steps: the 'E-step' finds the expectation of the missing data through sufficient statistics to model the density function; and the 'M-step' finds the best value for the parameter θ using complete-data ML techniques.

3.1.3. Bayesian estimation

In Bayesian estimation, information about unknown parameters is expressed in the form of a posterior probability distribution of all the unknown parameters. Bayes' Theorem for a continuous parameter θ and data points represented in a vector $\boldsymbol{y}$ is:

$$p(\theta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\theta)p(\theta)}{\int p(\boldsymbol{y}|\theta)p(\theta)d\theta}$$

This is often expressed in proportional terms by taking the denominator out as a constant of proportionality to give: $p(\theta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\theta)p(\theta)$ which gives rise to the Bayesian mantra: *posterior is proportional to the prior times the likelihood*

On the right hand side we have the likelihood PDF $p(\boldsymbol{y}|\theta)$ multiplied by the prior density function $p(\theta)$. This representation highlights the precise nature of Bayesian inference, and how it differs from frequentist inference. The frequentist approach is to work with a fixed (but unknown) property of a population (say parameter θ) and to randomly select sample data y. Repeated random samples will give different y and different sample-based estimates of θ, each denoted by $\widehat{\theta} = \hat{\theta}(y)$; a notational form that stresses estimates of

parameters are functions of data. The distribution of values of $\hat{\theta}$ that results from repeated application of the sampling process gives the sampling distribution of $\hat{\theta}$. The standard deviation of that sampling distribution of $\hat{\theta}$ is the standard error of $\hat{\theta}$. In summary, for frequentist inference, parameters are *fixed* and data are considered as *random*.

In contrast, the Bayesian model is founded on parameters being *random* variables determined by *fixed* observed data; the randomness of the model being summarised by the posterior density function $p(\theta|\boldsymbol{y})$. Bayesian procedures are centred on what we know or can say about θ in light of the data available for analysis. The Bayesian approach is sequential, in that the posterior findings can then be used as prior beliefs with new data. The computer software for Bayesian analysis uses the sample-based estimate $\hat{\theta}(y)$ as a helpful starting point in computing the posterior distribution for θ, but other than that, this estimate has no special place in Bayesian procedures. Indeed the roles of θ and $\hat{\theta}$ in Bayesian analysis are reversed in comparison to the frequentist approach, with θ being *random* as far as the analyst is concerned, i.e. parameters are *random* and data are considered as *fixed*. The uncertainty of the value of θ is ultimately only computed via simulation. This is one of the attractions of the Bayesian framework, as it provides a means to tackle complex problems that do not lend themselves to direct analytical solution. Analysis is by simulation, with pseudo-random draws made from the prior distribution. Reference to pseudo-random draws is necessary because computer generated random numbers follow a deterministic pattern that calls into question the truly random nature of the draws from the distribution. The pseudo-random label is used to signify the number generator only provides approximately independent draws. However, in all practical applications where a large number of samples is considered any function of the sampled values will be arbitrarily close to the corresponding feature of the distribution.

3.1.4. Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is a collection of techniques that generates the required pseudo-random draws from probability distributions. The essential idea in Monte Carlo methods that is applicable to Bayesian computation hinges on the *Monte Carlo principle:*

> anything we want to know about a random variable θ can be learned by sampling many times from $f(\theta)$, the density of θ. (Jackman, 2009: 133)

With the increasing availability of computer memory and speed of processing, the reference to ‘many times’ can readily be accommodated. A larger number of samples equates to greater precision in the estimates, a feature that makes the Monte Carlo method

simulation-consistent. The technique itself is not new, with examples of Monte Carlo methods being used well before the availability of computers. For example, Hall (1872) presented a short paper on an experimental determination of π, using a long run of 'Buffon's needle experiment' to estimate the value of π. The larger the sample, the better the approximation, with the probability determined through a long run of experimentation getting increasingly closer to the analytic solution that made use of calculus. See Appendix 3.1: Buffon's needle experiment for full details. Monte Carlo methods are equally useful whether or not the posterior distribution is of a standard form – indeed learning about non-standard functions of parameters is where its strength lies. This makes the method well-suited to problems in the social sciences where the posterior distributions tend to be non-standard and difficult to work with analytically. If random samples can be drawn from the prior density then Monte Carlo methods can generate simulation-consistent information about the prior. This technique shifts the problem from being mathematically complex to an easier problem of how to draw a large number of samples, and back to the Bayesian mantra: *the posterior being proportional to the prior times the likelihood*.

When no strong prior information is known about θ it is common practice to assume a uniform distribution over an appropriate range of values as a non-informative prior distribution. This avoids the introduction of bias that might result from imposing a structure that does not fit the data. It is also an entirely reasonable starting point from a Bayesian perspective, where analysis rests on the observed data with an assumption of prior ignorance.

In the standard linear regression model there are essentially three parameters to be estimated as one strives to generate samples from the distribution $p(\beta_0, \beta_1, \sigma_e^2|y)$. The default priors utilized in MLwiN when MCMC estimation is used are 'diffuse' or 'flat' for all parameters (Browne, 2012: 4) The default prior for fixed parameters in MLwiN is $p(\beta) \propto 1$. This *improper* uniform prior is equivalent to a *proper* Normal prior with a very large variance; i.e. an extremely diffuse distribution. It is an *improper* prior distribution because it does not integrate to 1 and is therefore not a true probability distribution, but the resultant posterior distribution will be a *proper* distribution. The default prior for scalar variances is, $p(\frac{1}{\sigma^2}) \sim \Gamma(\varepsilon, \varepsilon)$, where ε is very small. This *proper* prior is more or less equivalent to a Uniform prior for $\log(\sigma^2)$. When MCMC estimation is extended for multilevel models additional steps are added to the algorithm. The most notable addition is the level 2 variance matrix that is updated by drawing from its inverse-Wishart full conditional. The MLwiN default prior for variance matrices is $p(\Omega^{-1}) \sim \text{Wishart}_{\text{p}}(\text{p}, \text{p}, \widehat{\Omega})$, where p is the number of rows in the variance matrix and $\widehat{\Omega}$ is an estimate for the true value of Ω. The first parameter, which represents the sample size on which the prior belief is

based, is set to the smallest possible value so that this prior is only weakly informative. The estimate $\widehat{\Omega}$ will be the starting value of Ω (usually from the Iterative Generalised Least Squares (IGLS) estimation routine as cited above) and so this prior is essentially an *informative* prior. Indeed, starting values for each parameter are *needed* before MCMC estimation can be started. For example the ML estimates generated through IGLS procedure are used by MLwiN as good starting values for $\beta_0(0), \beta_1(0) and\ \sigma_e^2(0)$.

When prior beliefs are diffuse or uninformative as outlined above, the observed data likelihood will dominate the posterior distribution of the parameters – as intended within Bayesian inference. When prior information is available it will be incorporated into the posterior, but if priors are stated too precisely relative to the data, then the likelihood has less of a contribution and this could lead to the data being largely ignored. This is counter to intended Bayesian inference which is why there is such an emphasis on working with diffuse prior distributions. The ML estimates merely give a starting point for MCMC technique to be followed through and are not as such imposing a strong prior function.

As discussed above, the Monte Carlo methods let us learn about a random variable by sampling many times. The Markov Chain component of MCMC provides the random process to the method. A Markov process or Markov Chain is a random process or sequence of events that satisfies the following definition provided by Jackman (2009: 172):

> Let $\{\theta^t\}$ be a stochastic process, a collection of random variables indexed by time, t. If the stochastic process has the property that
> $\Pr\left[\theta^{(t+a)} = y \middle| \theta^s = x_s, s \leq t\right] = \Pr\left[\theta^{(t+a)} = y \middle| \theta^t = x_t\right], \forall\ a > 0$
> Then the process is said to possess the Markov property, and any such process is said to be a Markov process or Markov chain.

Essentially this definition stipulates conditional independence; whereby predictions on future outcomes are only dependent on the immediate past and do not depend on earlier states in the process. Applied to the sampling process in MCMC, a Markov Chain is a random process such that it visits locations in the parameter space with frequencies proportional to the probability of those locations as predicated by the posterior density distribution, $p(\theta|data)$. A Markov Chain with this property is said to be *ergodic:* a process in which every sequence or sizeable sample is equally representative of the whole. Although the samples are not strictly independent, the ergodic theorem or law of large numbers for Markov Chains means that averages $\bar{h} = \frac{1}{T}\sum_{t=1}^{T} h\left(\theta^{(t)}\right)$ taken from the Markov Chain output are simulation-consistent estimates of the posterior distribution. No matter where you start the Markov Chain, so long as it runs for long enough, it will eventually generate a random tour of the parameter space, proportionally visiting locations as befits the posterior distribution. This lead phase of the chain is known as the 'burn-in' time, iterations that are

discarded from the sample to allow the chain to settle down before sampling from the data-driven posterior distribution. In MLwiN the default burn-in period is 500 iterations, after which the chain will have converged to its equilibrium distribution. Diagnostic checks can confirm whether a longer burn-in is desirable, for instance in cases where there is still evidence of auto-correlation or indicators that the full sample space is not as yet being included in the chain. If we want several samples we can (i) repeat the process (ii) take consecutive samples after the burn-in time or (iii) extend the random walk through further iterations and record every l^{th}-step as a sample. The third method will result in approximately independent samples from θ if l is sufficiently large; the second method will not produce independent samples (Lebanon, 2006). However, Monte Carlo methods in conjunction with Markov Chains let us learn about a random variable even when the samples from the distribution are not strictly independent, but 'auto-correlated' with subsequent terms serially dependent. This was a feature of the technique that particularly interested Andrey Markov (1856-1922), more so that any application of his model, as he showed that independence was not necessary for large numbers.

Bayesian modelling in MLwiN uses a combination of two Markov Chain Monte Carlo (MCMC) procedures: Gibbs sampling and Metropolis-Hastings sampling. Gibbs sampling from a joint posterior distribution breaks the task into stages, sampling from conditional posterior distributions in turn to simulate new values for the parameters assuming current values for the other parameters are true. This process completely characterises the joint posterior function. Taking the same example as cited above for a standard linear regression model, the aim is to generate samples from the distribution $p(\beta_0, \beta_1, \sigma_e^2|y)$ to specify parameters of the distribution that will serve as 'priors' in subsequent iterations. The steps involve successively drawing from the conditional distributions until all parameters or sub-vectors are updated from their initial values determined by ML estimates (using IGLS). For the standard linear regression we have initial ML values for $\beta_0(0), \beta_1(0)$ and $\sigma_e^2(0)$, and set out to find $\beta_0(1), \beta_1(1)$ and $\sigma_e^2(1)$ from their respective conditional distributions that take account of the latest estimate for each parameter on subsequent steps:

$$\beta_0(1) \sim p\big(\beta_0|y, \beta_1(0), \sigma_e^2(0)\big)$$
$$\beta_1(1) \sim p\big(\beta_1|y, \beta_0(1), \sigma_e^2(0)\big)$$
$$\sigma_e^2(1) \sim p\big(\sigma_e^2|y, \beta_0(1), \beta_1(1)\big)$$

After a full set of parameters are drawn the process is repeated to obtain $\beta_0(t), \beta_1(t)$ and $\sigma_e^2(t)$ and so on. The sequence of samples $\theta(t): t = 0,1,2, \ldots$ forms a Markov Chain where every new value generated for a parameter only depends on its previous values

through the last value generated, $\theta(t-1)$. The Markov Chain has a stationary distribution equal to θ for large *t*.

Metropolis-Hastings sampling is a more general MCMC estimation method. MCMC estimation methods generate new values from a *proposal* distribution that determines how to choose a new parameter value given the current parameter value. The proposal distribution under Gibbs sampling is the posterior distribution, with each new value accepted as the estimate for the next iteration – a special case of the Metropolis-Hastings sampler where every proposed value is accepted. The general model either accepts the new estimate for the next iteration or rejects in favour of the current value for the next iteration. The plausibility of the proposed parameter value relative to the current value is assessed though an acceptance ratio as detailed below. In MLwiN (Browne, 2012), the Metropolis-Hastings sampler uses Normal proposal distributions centred at the current parameter value. To illustrate how the updating procedure works for the parameter β_0 at time t, $\beta_0(t)$, in the Normal variance components model we have three stages:

- Draw β_0^* from the Normal proposal distribution $\beta_0(t) \sim N(\beta_0(t-1), \sigma_p^2)$ where σ_p^2 is variance of the proposal distribution.
- Define acceptance ratio

$$r_t = \frac{p(\beta_0^*, \beta_1(t-1), \sigma_e^2(t-1)|y)}{p(\beta_0(t-1), \beta_1(t-1), \sigma_e^2(t-1)|y)}$$

  and let $a_t = \min(1, r_t)$ be the acceptance probability. A proposal distribution with small variance ($\sigma^2 \to 0$) will have strongly correlated consecutive samples and a high acceptance probability. Increased variance would reduce the correlation between consecutive samples but would also reduce the acceptance probability.
- Accept the proposal of β_0^* for $\beta_0(t)$ with probability a_t,
  otherwise let $\beta_0(t) = \beta_0(t-1)$, with no change to the parameter

The acceptance rate is inversely proportional to variance of the proposal distribution, a feature that presents a challenge when using Metropolis-Hastings algorithms in setting an appropriate proposal variance i.e. finding a 'good' value for σ_p^2 . Extremes on the continuum are to be avoided:

> Firstly choosing too large a proposal variance will mean that proposals are rarely accepted and this will induce a highly auto-correlated chain. Secondly choosing too small a proposal variance will mean that although we have a high acceptance rate the moves proposed are small and so it takes many iterations to explore the whole parameter space again inducing a highly auto-correlated chain. (Browne, 2012:18)

In most situations the missing values cannot readily be resolved analytically and so alternative methods are used to estimate the parameters. The Expectation-Maximisation (EM) algorithm is used for maximum likelihood estimation, and Markov Chain simulation methods are used in computing Bayesian estimates. Computer software has been developed to support both types of calculation that lead to the identification of suitable estimates for parameters of interest. For example in MLwiN, the Markov Chain Monte Carlo (MCMC) method can be used to estimate the parameters of the posterior distribution, where the Markov Chain's equilibrium distribution is used as a sampling point for the desired distribution. The challenge for analysts is to determine how many steps (iterations) are needed to converge to the stationary distribution that will provide those statistics with an acceptable margin of error. A range of diagnostic tools are available to assess the trajectories of the likelihood trace and parameter estimates, and both convergence and accuracy diagnostics. These will be discussed in the analysis section where actual values, graphical distributions and accuracy statistics can be used to illustrate the findings and the margin of error reported through MCMC methods.

3.1.5. Single imputation

Inferences about the parameters are made by looking at the features of the posterior distribution. However, inferences based on a single imputation tend to underestimate the inherent variation and as Rubin (1987) points out, will be too 'sharp', since that extra variability due to the unknown missing values is not taken into account. For example in a simple random sample with no covariates, complete data estimates can follow from the statement that $(\bar{y} - \bar{Y})$ is normally distributed with mean zero and variance $s^2\left(\frac{1}{n} - \frac{1}{N}\right)$, where sample variance is based on a finite sample with:

$$n = sample\ size$$
$$\bar{y} = sample\ mean$$
$$s^2 = sample\ variance$$
$$N = population\ size$$
$$\bar{Y} = population\ mean$$

Because there are missing data in the sample due to non-response, there will be say n_1 observed values within the sample of n, that have a sample mean and variance of $\bar{y}_1$ and s_1^2. Standard inferences for $\bar{Y}$ are therefore based on the statement that $(\bar{y}_1 - \bar{Y})$ is approximately normally distributed with mean zero and variance

$$s_1^2\left(\frac{1}{n_1}-\frac{1}{N}\right)$$

But if the best prediction values of the missing *Y* are imputed using *mean substitution* then each missing *Y* is the mean of the observed sample mean, $\bar{y}_1$. The overall mean for the *n* values is $\bar{y}_1$, with n_1 observed values having mean $\bar{y}_1$ and the remaining $(n-n_1)$ imputed values each being $\bar{y}_1$. The sample variance for all *n* values is therefore

$$s_1^2\left(\frac{n_1-1}{n-1}\right)$$

given that the imputed values will have zero variance, leaving all of the variance as based on the n_1 observed values. This can be shown through the following calculations:

With $s^2=\frac{\sum_1^n(y_i-\bar{y})^2}{n-1}$ and $s_1^2=\frac{\sum_1^{n_1}(y_i-\bar{y}_1)^2}{n_1-1}$, giving

$$\sum_1^{n_1}(y_i-\bar{y}_1)^2=(n_1-1)s_1^2$$

Since $\sum_1^n(y_i-\bar{y})^2$ reduces to $\sum_1^{n_1}(y_i-\bar{y}_1)^2$ plus the zero values for the remaining $(n-n_1)$ imputed values that are equal to $\bar{y}_1$, the variance calculation becomes:

$$s^2=\frac{\sum_1^{n_1}(y_i-\bar{y}_1)^2}{n-1}=\frac{s_1^2(n_1-1)}{n-1}=s_1^2\left(\frac{n_1-1}{n-1}\right)$$

as above.

The estimate of population variance of *Y*, when based on these best prediction values for each missing *Y*, is too small by the factor $\left(\frac{n_1-1}{n-1}\right)$.

Rubin concludes that the variance of $(\bar{y}-\bar{Y})$, based on the data set being completed by mean substitution imputation, is therefore $s_1^2\left(\frac{1}{n}-\frac{1}{N}\right)\left(\frac{n_1-1}{n-1}\right)$, which in comparison to the actual variance of $\bar{y}_1$ is too small by a factor of $\left(\frac{n_1}{n}\right)^2$, for large values of n_1 and $\frac{N}{n_1}$ (Rubin, 1987: 14). Using a single imputation is not recommended as it will not lead to correctly centred inferences or interval estimates of the population mean $\bar{Y}$. The recommendation is therefore to consider multiple imputations.

3.1.6. Multiple imputation

The use of multiple imputation (MI) is merely an extension of single imputation, with a prime goal of reducing the variance in parameter estimates that is due to the missing data values. The use of MI was first proposed by Rubin in 1978 and further developed by him in the context of large sample surveys (Rubin, 1987). The basic concept is to have a mechanism that will generate a set of plausible values for each missing value. Rubin

discusses the use of m imputed sets, where $m > 1$, with resultant statistics combined using techniques to give parameter estimates and standard errors that address the earlier concerns of estimates being too 'sharp'. This is achieved by factoring in an additional variance component that takes into account the uncertainty due to the missing data values; the total variance combines two components to reflect the *within-imputation* and the *between-imputation* variance. The formulae used by Rubin (1987) and Schafer (1999), referred to as Rubin's Rules, are detailed in Appendix 3.2: Combining multiple imputations using Rubin's rules.

Rubin goes on to show that only a small set of replacement values is necessary, with little benefit accrued from increasing the size of that set beyond five. For example, with up to 30% missing information, 94% efficiency in estimation is achieved with five imputations ($m = 5$); for the same level of missing information with ten imputations, the efficiency only increases to 97%. Any increase in the number of imputations carries a significant workload in generating and combining the resultant statistics, hence Rubin's argument that just three to five imputations are sufficient to obtain accurate results in any MI process. The justification for this conclusion is as reproduced in Appendix 3.3: Efficiency of using *m* imputations. An illustration of computational processes is included in Appendix 3.4

3.2. Survey design

Three key strands emerge from the survey design of TIMSS. First the sampling design makes use of the hierarchical structure of data in education as it seeks to secure representative samples of students across and within nations. The use of hierarchically structured data in the surveys presents potential difficulties for analysts who must take account of that sample design when selecting methods and software to analyse and interpret the findings. Second, analysts must take into account the matrix-sample design used to assign tasks and questionnaires to those included in the sample. The matrix-sampling technique presents incomplete data at the individual level. The missing data needs to be handled appropriately, with imputation techniques used to generate complete-data sets that thereafter allow analyses using standard approaches. The final issue related to survey design concerns the structure of data gathered through questionnaires. At the data collection stage, many questions are presented with optional responses on either an ordinal or polytomous scale; subsequent analyses rely on combinations of such responses. To avoid over-simplification and potential loss of data, analysts can consider a re-scaling of these data as discussed below.

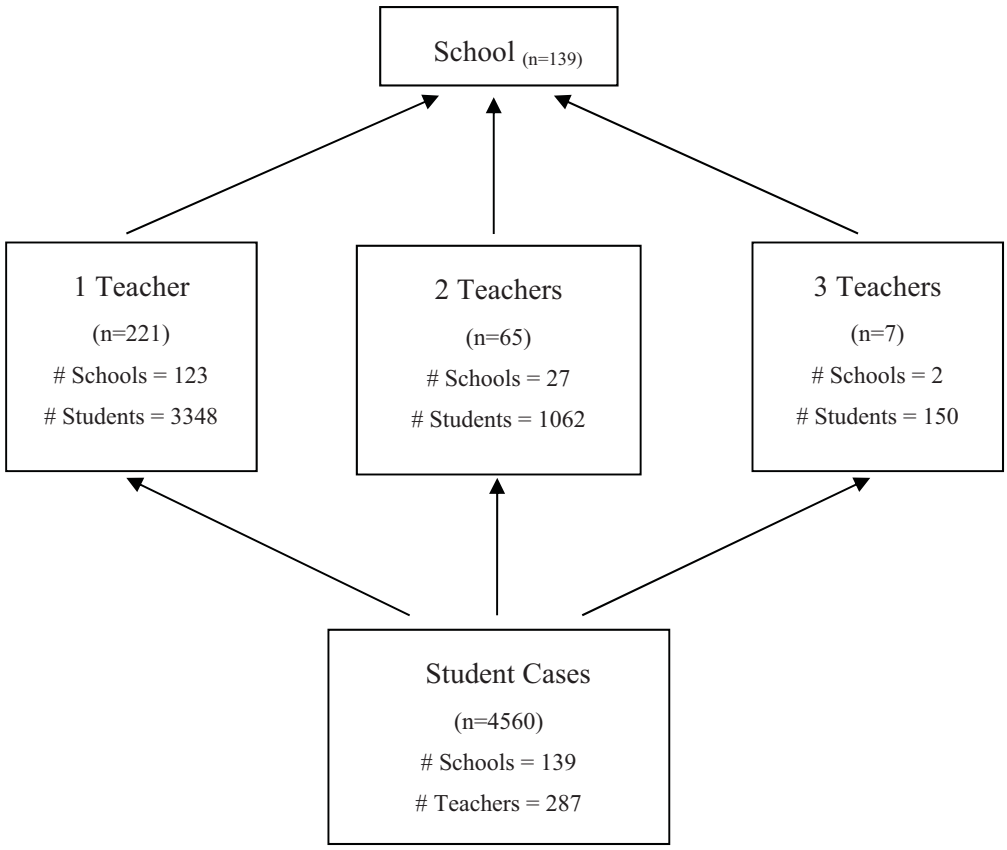
3.2.1. Clustered data and hierarchical models

The basic configuration of the target populations from which the samples are drawn is highly structured with students identified for inclusion by country, stage, school and class. Within each country there are two target populations represented by age and stage, with samples drawn to represent the national equivalent of students in G4 and G8. These stages reflect the number of years of formal education and also satisfy the requirement for students' mean age at time of testing to be at least 9.5 and 13.5 years respectively.

To obtain accurate and representative samples, a two-stage stratified sampling procedure is deployed whereby schools are randomly sampled with probability proportional to size as the first stage, and then usually one intact class, occasionally two or three classes of students, are sampled at the second stage. All students in selected classes are assessed.

At first sight the structure of the TIMSS data sets appears to be a nested hierarchical relationship of units, with students identified as the atomic unit. On closer inspection some students are associated with more than one teacher and as such are included as multiple cases. For instance in $TIMSS_{2007}$ at G4 there are 4560 student cases in 139 schools with the structure and composition as displayed in Figure 3.2-1.

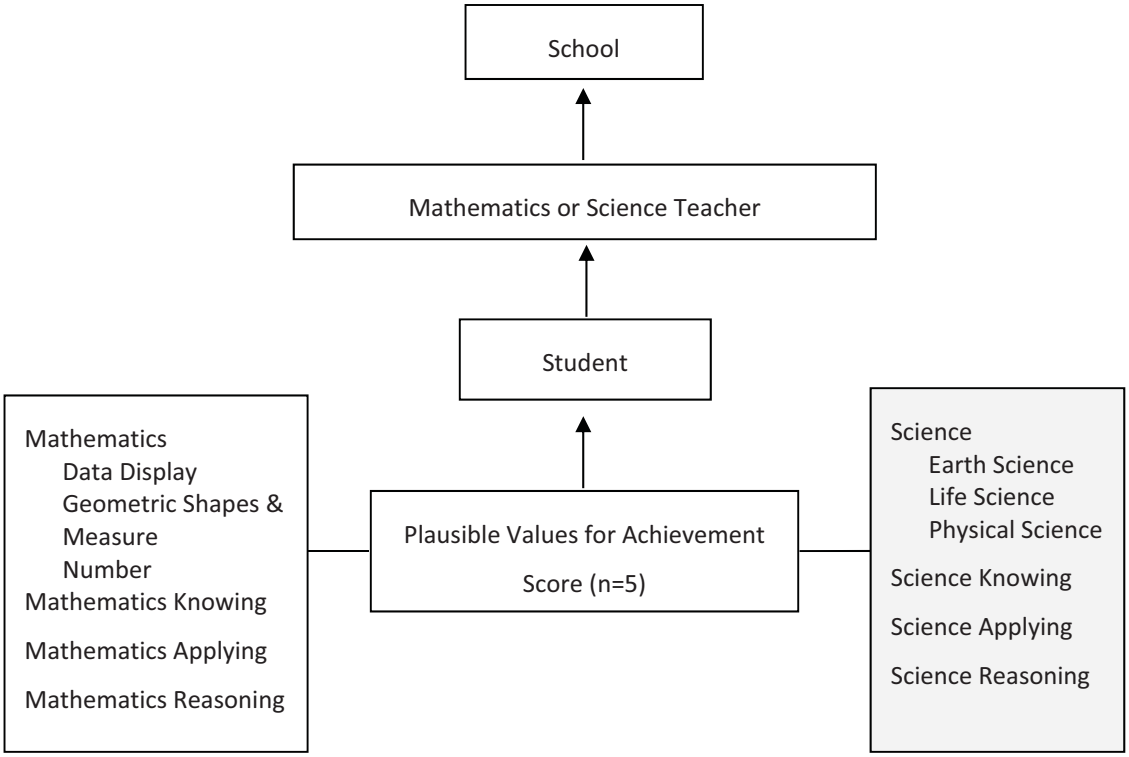
Figure 3.2-1: All G4 Student Cases and Teachers (Mathematics and Science)



Students have multiple teachers because of a separate input for mathematics and science, but also because of a job share or some other arrangement that results in having more than one teacher for mathematics and/or science. For the analyses of $TIMSS_{2007}$ the data sets were split by discipline. The duplicate and redundant student cases were dropped to generate a reduced data set that had a simple hierarchical structure with one teacher aligned with each student. The full details of this structure for mathematics and science instruction at G4 is presented in Appendix 3.4.

The primary analysis as set out in 3.3.2 takes plausible value as the atomic unit, creating a four-level hierarchical structure of repeated scores of achievement within student cohorts, which are then nested within teachers and within schools, as illustrated in Figure 3.2-2.

Figure 3.2-2: General hierarchical structure of data with plausible values at lowest level



For each of the domains in mathematics and science there are five plausible values (PVs) calculated per student. In my primary analysis I propose to explore the merits of incorporating those PVs within a multilevel model as a first level, taking PV as the atomic unit. However, for the purposes of the initial explorations, exploratory data analysis, and development of student achievement models for mathematics and science, I will use the

weighted average of the plausible values, applying the total student weight (TOTWGT[1]) for those within-country analyses.

Models of achievement scores are built up to include a range of explanatory variables at student, teacher and school levels of analysis. The findings reported in Bidwell & Kasarda (1980), Hattie (2009), Hopkins (2000), Leithwood *et al.* (2008), and Yair (2000) all contribute to and influence the conceptual model and subsequent analyses of data at student, teacher and school levels of $TIMSS_{2007.}$ Yair's conceptual model of 'Engagement and Alienation from Instruction' (Yair, 2000: 250) separates background characteristics and external influences from the school effects of curriculum and instruction. A similar breakdown is pursued within the current $TIMSS_{2007}$ data set that includes background data from students, teachers and schools. The potential associations as outlined in theoretical frameworks of Chapter 2 are explored through relevant variables in EDA of questionnaire responses in Chapter 4. Distinctions are drawn between home and school characteristics as control variables, as well as acknowledging contributions related to personal and academic experiences, the latter reflecting on 'traditional' and 'reform' pedagogies as discussed in Chapter 2.

Given the split between the responses from the teachers of mathematics and science the selection criteria were different for each discipline. A complete case data set was achieved by filtering out the missing cases for any of the explanatory variables included in final models; this decision allowed analyses to be completed on the same cases within the nested structures and to evaluate contribution of explanatory variables according to the achievement model for mathematics or science respectively.

The resulting student sample presents a complex structure with students clustered within classes, within schools, within countries. This clustered, hierarchical structure must be taken into consideration when analysing the data. A major consideration when analysing clustered data is that the observations within a cluster cannot be assumed to be independent. Students in the *same class* will generally have similar learning and teaching experiences; students in the *same school* will have similar characteristics as based on geographical location, school composition in terms of socio-economic status and school culture or ethos; students in the *same country* will have similar experiences to one another in terms of cultural and political influences, and policy initiatives pursed in education. Students in those clusters will be more like each other than like other students in the target population, so sampling one intact class of 25 pupils will provide the researcher with less information per student than a similarly sized random sample of students drawn from across all students in the target

[1] A discussion of weights is developed in section 3.2.3

population. The degree of similarity within each cluster, be it at class, school or country level, can be measured by the intra-cluster correlation coefficient (ICC), which acknowledges and quantifies the similarities of characteristics and experiences of students selected for inclusion in the study. Any disregard for the special characteristics of the sample design will tend to underestimate the true sampling variability, leading to confidence intervals that are too small, biased estimates and incorrect interpretation of associations between variables (Beaton, 1987; Beaton and Zwick, 1992).

Clustered data imply a hierarchical structure that can be analysed at various 'levels'. In $TIMSS_{2007}$, the data structure permits several levels of analysis (student-, teacher-, school- and national-levels are the obvious starting points) and presents considerable opportunities to explore the interrelationships among variables at any level. These analyses, generically called multilevel analyses, will take student's achievement score as the outcome measure at the lowest level of data in the study. The multilevel data analyses aim to assess the effect of student's actions and classroom experiences on achievement measures. Similarly, an analysis of teachers' practices, and school or national policies, can be undertaken to determine whether variables from higher levels of the data structure serve as moderators of student's achievement, i.e. providing an explanation for variances between student's achievements. Multilevel models are designed to analyse variables from different levels simultaneously, taking due account of any intra-cluster correlations that impact on estimates of variability and resultant interpretation of associations between variables. Due consideration of these issues is acknowledged in 3.3.2, where random-effects regression models are outlined as the primary method of analysis.

3.2.2. Derived variables

Throughout the analyses a number of derived variables were computed to combine existing variables into meaningful and manageable groupings. These computations ranged from simple averages of related variables and indices, to the derivation of new components using Principal Component Analysis (PCA). The former are much as computed by IEA in their published derived variables such as used for Students' Perception of Being Safe in School (SPBSS coded as ASDGPBSS). This derived variable (noting 'D' within code) was an index code of High, Medium and Low in response to five statements, where High corresponded to 'No' on all statements and Low was attributed to having three or more 'Yes' responses to the five statements:

- Something of mine was stolen (AS4GSTOL)
- I was hit or hurt by other student(s) (e.g., shoving, hitting, kicking) (AS4GHURT)
- I was made to do things that I didn't want to do by other students (AS4GMADE)

- I was made fun of or called names (AS4GMFUN)
- I was left out of activities by other students (AS4GLEFT) (Foy & Olson, 2009)

Calculations such as those and any derived variables identified through PCA were computed using SPSS. A worked example of how PCA was carried out is included in Chapter 4, where eight 'Home Resource' variables were reduced to two components plus two separate variables retained in their own right. Kaiser's Eigenvalue-one criterion and Catell's Scree graph guided the selection of variables as recommended in Tinsley & Tinsley (1987). The principal axis method was used to extract components, followed by a varimax (orthogonal) rotation to generate uncorrelated components suitable for inclusion as independent variables in subsequent analyses; Varimax rotation is commonly used in the social sciences. Interpretation of findings and composition of components was guided by a cut point of 0.4 on factor loadings. Any factor loading lower than 0.4 was dropped and remaining factors were combined with equal weighting. This resulted in *factor-based* scores being used as opposed to factor scores – justification and wider discussion of these processes are included in Chapter 4 alongside a worked example.

Where several variables contributed to a component, the scoring mechanism comprised an extended categorical scale. A rescaling mechanism was pursued where there were grounds to consider an underlying continuous metric in line with methods promoted by Snell (1964). The theory of this is outlined in 3.2.6 with an illustration of practice in Chapter 4.1.4 (Out of school interest and experiences).

3.2.3. EDA using weighted data

The two stage stratified cluster sampling design outlined in 3.2.1, started with selection of schools, systematically sampled (random start, fixed interval) with probability proportional to size (PPS). The effect of this strategy is that an individual's probability of selection depends on which stratum they belong to, with the first stratum being size of school; intact classes are then randomly sampled from the desired target population. All schools in the sampling frame were sorted by measure of size (MOS), an explicit stratum, and then sorted by any implicit strata agreed between the country and IEA. The implicit strata in Scotland were school grade (G4 & G8), urbanization (large urban area, other urban area, accessible small town, remote small town, accessible rural area, remote rural area), and school deprivation index (low, middle, high, unknown), making up a total of 33 implicit strata. School-level exclusions consisted of very small schools (MOS <6), special schools and Gaelic schools. Within schools, exclusions extended to those with special educational needs and non-native language speakers. Given those strata and sampling restrictions the

student sample figures for Scotland and comparator countries are as noted in Table 3.2-1and Table 3.2-2 for G4 and G8 respectively.

Table 3.2-1: Population and Sample Sizes (G4)

Country	Population		Sample			Average Age at Time of Testing
	Schools	Students	Schools	Students	Estimated population	
Denmark	1,800	67,200	137	3,500	59,300	11.0
England	15,300	608,100	143	4,300	578,600	10.2
Netherlands	6,600	186,900	141	3,300	168,100	10.2
New Zealand	1,800	56,400	220	4,900	55,100	10.0
Norway	2,200	60,800	145	4,100	58,000	9.8
Scotland	1,900	58,100	139	3,900	55,000	9.8
Sweden	3,600	112,100	155	4,700	94,000	10.8

Table 3.2-2: Population and Sample Sizes (G8)

Country	Population		Sample			Average Age at Time of Testing
	Schools	Students	Schools	Students	Estimated population	
England	3,900	636,700	137	4,000	583,200	14.2
Norway	1,100	62,300	139	4,600	58,800	13.8
Scotland	400	64,800	129	4,100	59,300	13.7
Sweden	1,500	125,500	159	5,200	117,300	14.8

For any international comparisons it is worth noting that the target population in Scotland is amongst one of the youngest to be tested at both stages. That is on top of the fact the students have effectively had an 'additional' year of schooling, with the sample drawn from Primary 5 and Secondary 2 stages of the Scottish system for the respective G4 and G8 cohorts; there are seven years of study in Scottish Primary schools before transfer to Secondary schools.

An important feature of survey analysis is that samples are carefully drawn from the target population and the sample design needs to be taken into account in order for those sample data to accurately reflect population characteristics. Sampling weights are applied to take account of selection, stratification, non-response and any disproportionate sampling of clusters. The samples and estimated populations reported in the tables above have been calculated using sample weights that were calculated to account for the selection of each student, teacher and school participating in $TIMSS_{2007}$. Sampling weights are needed to account for the fact that units are selected with differing probabilities. Taking student as the unit level, every student's response is adjusted to reflect the actual proportional presence in the target population. For each student an overall sampling weight was calculated as the product of three components – a *school weight* calculated from the probability of selecting a school from the target population (the inverse of the probability of selection); a class or

teacher weight to reflect the probability of that class being chosen within the school; and finally a *student weight* for the probability of being selected in the class. In most cases the *student weight* will equal 1 when an intact class is selected i.e. the student is selected with certainty. This basic weight is then adjusted as necessary to respond to other sampling factors such as non-response, where schools, classes or students are withdrawn and *a priori* replacement units are substituted to make up the required sample. Fuller details on the processes and sampling criteria that had to be met on first wave and replacement samples are well-documented in the $TIMSS_{2007}$ User Guide (Foy & Olson, 2009).

The $TIMSS_{2007}$ data includes several different weights for analysts to choose from dependent on type of analysis being completed. These are set out in Table 3.2-3.

Table 3.2-3: Weights used in $TIMSS_{2007}$

Weight	Description
TOTWGT	Total student weight – sums to the national population
SENWGT	Student senate weight – sums to 500 in each country
HOUWGT	Student house weight – sums to the student sample size in each country[2]
SCHWGT	School weight
TCHWGT	Overall teacher weight
MATWGT	Mathematics teacher weight
SCIWGT	Science teacher weight

The sum of the sampling weights of all students in a country is an estimate of the size of the target population. Whenever student population estimates are wanted, the student sampling weight TOTWGT (Total student weight) must be used to inflate the sample size to the size of the population. When comparisons are made across countries, using TOTWGT would result in reporting on a disproportionate number of students if the countries had different population sizes, because TOTWGT sums to the target population size of each country. An alternative weighting factor is SENWGT (Student senate weight), a transformation of TOTWGT that results in a weighted sample size of 500 in each country. Direct comparisons will therefore make sense, with each country treated equally, regardless of size differences across the target populations of G4 or G8. Student's HOUWGT (house weight), another transformation of TOTWGT, ensures that the weighted sample corresponds to the actual sample size in each country. If school- or teacher-level analyses are desired then the appropriate weighting variable (SCHWGT, TCHWGT, MATWGT, or SCIWGT) should be applied prior to higher levels of analyses.

All of the above is directly applicable for the EDA where TOTWGT was used to

[2] House and Senate weight are so called because of the function in the US Congress, where the number of representatives in the House of Representatives is based on each state's population size, whereas the Senate gives two seats to each state regardless of population (Rutkowski *et al.*, 2010)

weight the SPSS data for any descriptive analysis of the data. Faiella (2010) amongst others notes there is broad agreement over using survey weights in descriptive analyses but less clarity or consensus on automatically extending that expectation when studying associations among survey variables – an 'analytic use of sample surveys'. Binder and Roberts (2003) suggest that when the analyst is interested in the association between predictors and response variable the sample design is assumed to be ignorable, as discussed by Rubin (1976). The Bayesian approach suggested by Rubin is an example of a model-based analysis that relates closely to the current study using $TIMSS_{2007}$ data. Alexander (1987) comments on the theoretical debate over whether survey data weights need to be used at all, especially where analysts are making model-based inferences. One strategy to accommodate weights in regression models is to include the variables on which the sampling framework is determined as predictors within the model, thereby controlling by proxy for the weights. However, Alexander rightly concludes:

> proponents of weighting would assert that no model will include all the relevant variables, and that few analysts will wish to include in their model all the geographic and operational variables which determine sampling rates. It is difficult to object in principle with the goal of correctly modelling all relevant variables, including the variables relating to sampling. However, the theoretical and empirical task of deriving, fitting, and validating such models seem formidable for many complex national demographic surveys. Alexander (1987: 188)

The version of MLwiN I used did not directly support inclusion of sample weights and although consideration was given to include 'school size' as a predictor variable, since school size (MOS) was the main criterion in sampling framework, there were another 33 implicit strata at G4 and 29 implicit strata at G8; this made the task unrealistic and I therefore followed Alexander's lead and focused on a model-based analysis using Bayesian techniques to identify associations between predictor variables and achievement scores reported in $TIMSS_{2007}$.

3.2.4. Matrix sample design

A key feature of $TIMSS_{2007}$ is its use of a matrix sample design when allocating research instruments to the selected respondents. The matrix sample design provides many benefits to those managing this education survey. As with the origins of multiple matrix sample design, stemming from research on educational testing by Turnbull, Ebel and Lord in the 1950s, the attractions for researchers include the reduced demand on respondents and reduced costs in running the survey. Shoemaker (1973) summarised the statistical methodology involved and highlighted processes and guidelines for implementing the technique (cited in Gonzalez and Eltinge, 2007). The prime purpose for TIMSS is to

generate comparative data on trends in achievement in the context of different educational systems, settings and practices without placing excessive burdens on schools or students; it is not seeking to report on individual students but rather to comment on institutional and national profiles. A further driver behind using a matrix sample design is that it is infeasible and unreasonable for individual students to be asked all possible questions on a range of topics. The institutional and national picture that is sought can readily be achieved through a matrix sampling approach that solicits only a few responses from each student but maintains a wide coverage of content when responses are aggregated across all students. The methods used to summarise the data are detailed in section 3.2.5 below.

The matrix sample design for the cognitive items at G4 and G8 of $TIMSS_{2007}$ comprises 14 blocks for each of mathematics and science. At fourth and eighth grades, half of the blocks contain secure items from $TIMSS_{2003}$ (to measure trends) and the other half are newly developed items for $TIMSS_{2007}$. These blocks are combined, as shown in Table 3.2-4 to form Student Achievement Booklets where each student is presented with one booklet that covers a mix of Mathematics and Science items as well as a mix of 'trend' and 'new' items. Broad coverage of the cognitive domains is included in each of the student achievement booklets. The booklets provide coverage of the cognitive framework of $TIMSS_{2007}$, with 353 items at fourth grade and 429 items at eighth grade, but the matrix sample design means each student is only exposed to an average of 55 and 68 items in the respective stages.

Table 3.2-4: Composition of student achievement booklets

Student Achievement Booklet	Assessment Blocks			
	Part 1		Part 2	
Booklet 1	M01	M02	S01	S02
Booklet 2	S02	S03	M02	M03
Booklet 3	M03	M04	S03	S04
Booklet 4	S04	S05	M04	M05
Booklet 5	M05	M06	S05	S06
Booklet 6	S06	S07	M06	M07
Booklet 7	M07	M08	S07	S08
Booklet 8	S08	S09	M08	M09
Booklet 9	M09	M10	S09	S10
Booklet 10	S10	S11	M10	M11
Booklet 11	M11	M12	S11	S12
Booklet 12	S12	S13	M12	M13
Booklet 13	M13	M14	S13	S14
Booklet 14	S14	S01	M14	M01

Given the random allocation of booklet-to-student the unassessed cognitive items are deemed to be missing completely at random (MCAR).

3.2.5. Missing data and multiple imputation

There are two types of missing data in $TIMSS_{2007}$. First there are data missing as a result of the matrix-sample survey design, and second, there are missing responses to the background questionnaires requested of students, teachers and head teachers.

The first example of 'missingness' comes as a result of the planned design to have an incomplete data set as outlined above i.e. missingness is by design and as such is by random selection. In $TIMSS_{2007}$ and other similar educational studies, a matrix survey design is implemented with students randomly assigned test item booklets that only cover part of the cognitive domain. Since the students were randomly selected from within their peer group in the stratified sample of schools, the probability of a missing response is independent of all the measured and unmeasured characteristics of the students, making the missing cognitive items MCAR.

The second type of missing data is where there are cases of unit and item non-response in the background questionnaires. Responses here may be MCAR, missing at random (MAR) or missing not at random (MNAR). Any missing data will need to be reviewed and evaluated separately for each case of missingness. The way missing background data are handled, in terms of listwise or pairwise deletion, or by 'filling in' through an appropriate imputation strategy, will be justified in light of the mechanism of missing data.

Cognitive items

The cognitive items of $TIMSS_{2007}$ are the elements that are affected by the matrix-sampling methodology. In practice, students are exposed to only a selection of cognitive items, with others recorded as 'missing'. Instead of first computing estimates of missing responses and then aggregating these to estimate population parameters, the adopted approach uses all available data, students' responses to the cognitive items they were administered together with background data, to estimate directly the characteristics of student populations and subpopulations. To summarise the cognitive responses $TIMSS_{2007}$ utilises Item Response Theory (IRT) scaling to describe student achievement in the assessments of mathematics and science curricula; a variant on methods put forward by Sinharay *et al.* (2001) in section 3.1, and in line with developments promoted and discussed by Lord (1977), Hambleton and Cook (1977) and Mislevy *et al*. (1992b). The background data are used to strengthen the reliability of student scores, a process known as 'conditioning', taking account of background information that is relevant to and impacts on educational practices (Foy et al., 2008). In the IEA's primary analysis of $TIMSS_{2007}$, a principal

component analysis (PCA) is used to reduce the wide range of background variables in order to identify a manageable number for inclusion in the marginal analysis. Typically around 90 per cent of the variance in the data is accounted for by the selected components. These principal components are the conditioning variables in the marginal analyses that will result in consistent estimates of population characteristics. The proficiency for any student (θ_k), a required parameter for IRT models, is a randomly selected value from such a conditional distribution.

There are three different types of cognitive items that need to be modelled: (1) simple dichotomous responses scored correct or incorrect; (2) multiple choice items that are scored correct or incorrect; and (3) polytomous response tasks where there are more than two response options, e.g. with partial credit given for the response. To cater for the different types of item and scoring procedures used across the survey, three distinct IRT models are considered to generate the scaling distribution that provides the mechanism to identify proficiency scores for each student in mathematics, science and each of the sub-scales described earlier – i.e. the probability of the student correctly responding to any particular test item. The basic structure of a latent trait model for a simple dichotomous item describes the probability of a particular response based on two characteristics that describe the item's difficulty and its power of discrimination. When dealing with multiple choice items, a third characteristic is taken into consideration, to reflect the probability of getting a correct response by chance. The numbers of characteristics that influence the probability of a student correctly responding to an item give the generic names of 2-parameter and 3-parameter logistic models (2PL and 3PL). For completeness, the 1-parameter logistic model has only one characteristic that influences the resultant probability, namely item difficulty. This special case, known as the Rasch model, is where all items are deemed to be equally discerning, with easy items discriminating between low-ability students and hard items discriminating between high-ability students; discriminating at different positions on the underlying continuous ability scale, θ.

As the name suggests, each model uses a logistic function to relate student ability and item parameters to the probability of correctly responding to an item. The probability of a student with proficiency θ_k responding correctly to an item x_i $(i = 1,2,3, \dots)$ is given by:

$$P(x_i = 1|\theta_k, a_i, b_i, c_i) = c_i + \frac{1 - c_i}{1 + \exp(-D.a_i.(\theta_k - b_i)}$$

Where

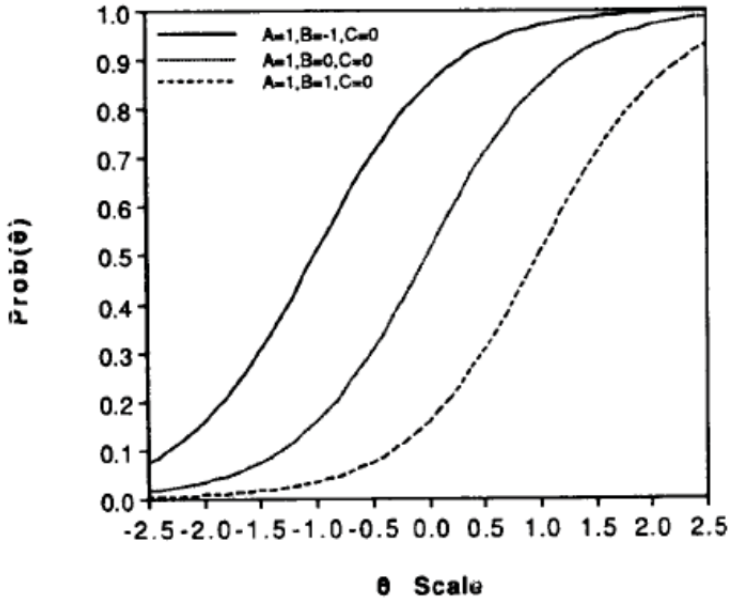
x_i is the response to item *i* (1 for correct, 0 for incorrect)
θ_k is the proficiency of student on the ability scale (higher proficiency has a greater probability of responding correctly)

a_i item discriminating power (slope parameter of item *i*)
b_i item difficulty (location parameter of item *i*)
c_i item chance, the lower asymptote of the item characteristic curve (ICC) that raises the probability of very low proficiency selecting a correct response (e.g. probability of being correct by chance in multiple choice items)
D an arbitrary scaling constant typically set to 1.7 to approximate results from the normal ogive model.

(Foy, Galia and Li, 2008: 226)

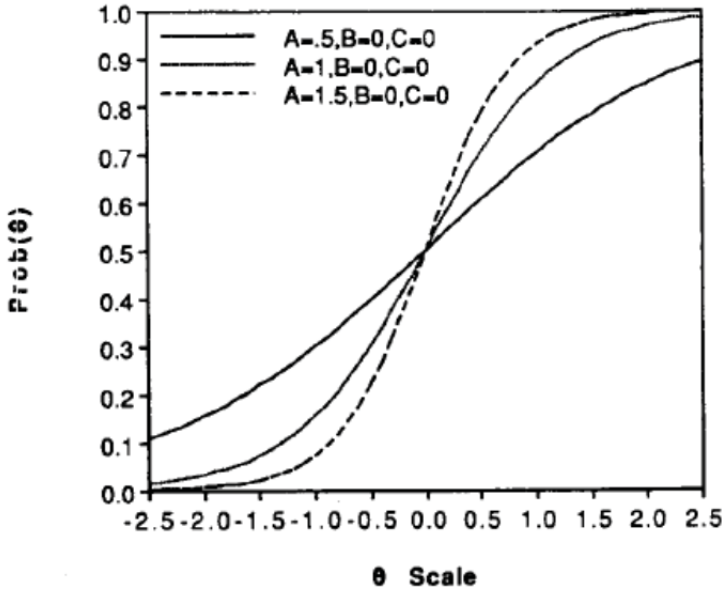
All models have an item difficulty parameter (b_i), which is the point of inflection on the ability scale (Figure 3.2-3: ICC graph). This determines the location of the parameter for any item, with the point of inflection for easy items positioned to the left of centre on the standardised proficiency scale (when $b_i < 0$) and to the right for harder items ($b_i > 0$).

Figure 3.2-3: ICC derived by varying parameter b_i in 3PL model (Harris, 1989: 159)



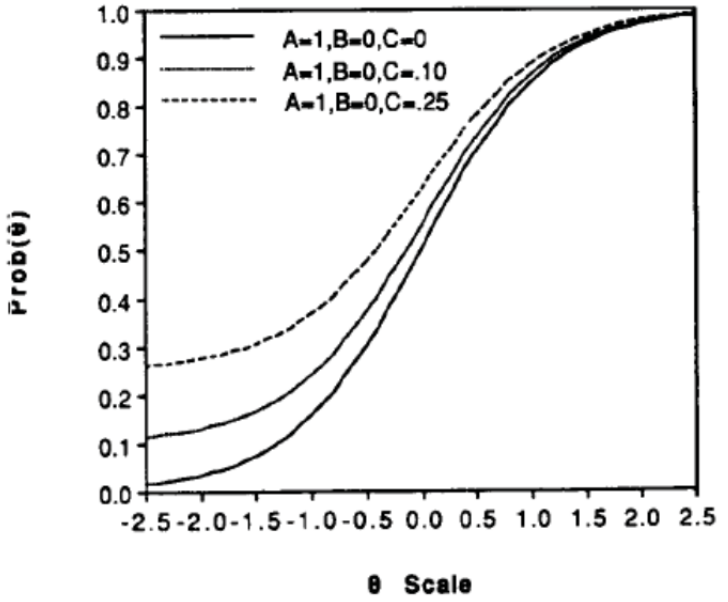
The 2PL curves for dichotomously coded constructed response can discriminate among students as shown in Figure 3.2-4 where the slopes of the curve vary. The inclusion of the a_i parameter allows the slopes of the curves to vary, indicating how discriminating the item is; the steeper the slope of the curve at the point of inflection, the more discerning the item.

Figure 3.2-4: ICC derived by varying parameter a_i in 3PL model (Harris, 1989: 159)



A 3PL model is utilised where multiple response items, each scored as correct or incorrect, are used to condition the proficiency scores generated through the model. The basic structure of the 3PL model takes account of a correct response by chance, including the c_i parameter to quantify the probability of correctly guessing from the multiple options. The 2PL and Rasch models assume no guessing or element of chance in the constructed response, with $c_i = 0$ in the general model provided above. As shown in Figure 3.2-5, as c_i increases to 0.25 (C=.25) the lower asymptote rises so that even those students with a low proficiency score are modelled to have an increased probability of success on the item.

Figure 3.2-5: ICC derived by varying parameter c_i in 3PL model (Harris, 1989: 159)



A third type of IRT model is used for polytomous response tasks where extended responses are scored with partial credit of 0, 1 or 2. Foy, Galia and Li (2008) present a generalised partial credit model that gives the probability a student with proficiency θ_k will have a response x_i for the i^{th} item, scored in the l^{th} of m_i ordered categories – in this case three ordered categories where the scoring is 0, 1 or 2.

$$P\left(x_i = l \middle| \theta_k, a_i, b_i, d_{i,1}, \dots, d_{i,m_i-1}\right) = \frac{\exp\left(\sum_{v=0}^{l} D.a_i.\left(\theta_k - b_i + d_{i,v}\right)\right.}{\sum_{g=0}^{m_i-1} \exp\left(\sum_{v=0}^{g} D.a_i.\left(\theta_k - b_i + d_{i,v}\right)\right)}$$

Where

m_i	is the number of response categories for item *i* (here $m_i = 3$ categories)
x_i	is the response to item *i* , ranging from 0 to $m_i - 1$ (in this case 0 to 2)
θ_k	is the proficiency of student on the ability scale (higher proficiency has a greater probability of responding correctly)
a_i	item discriminating power (slope parameter of item *i*)
b_i	item difficulty (location parameter of item *i*)
$d_{i,l}$	is the category *l* threshold parameter ($l = 0, \dots m_i - 1$)
D	an arbitrary scaling constant typically set to 1.7 to approximate results from the normal ogive model as above.

IRT modelling is based solely on the conditional parameters, i.e. the student's proficiency score and the parameters of the item. Item response probabilities are unaffected by other items present in the test or any other student characteristics or data collection conditions – this assumption of 'conditional independence' underpins IRT modelling.

The imputations for like students – students with similar response patterns and background characteristics in the sampled population – are drawn from the IRT scaling distributions that describe the population. This approach is known as *plausible values* methodology (Foy, Galia and Li, 2008; Mislevy, 1992b) where each imputation is a *plausible* score for that student, based on their cognitive responses and background characteristics.

Plausible values through multiple imputation

The emphasis throughout this study is on generating consistent estimates of major population characteristics directly from item responses (Dempster *et al.* 1977; Mislevy, 1984, 1985). Mislevy (1992a: 135) highlights the significant turning point for designing and analysing large-scale educational assessments, where upon population characteristics could be estimated accurately without first securing accurate estimates for individual students (Lord,1962; Sirotnik and Wellington, 1977). This approach is used in $TIMSS_{2007}$, where plausible values methodology provides a proficiency score in mathematics, science and each of the sub-scales described earlier. These proficiency scores can be used for the purposes of

secondary data analysis and to provide practitioners, policymakers, and the general public with meaningful measures of student achievement and progress. The accuracy with which any one plausible value can summarise an individual's score is clearly restricted by the model, given the imputed score is based on limited information and will almost certainly include an error. To minimise these errors a multiple imputation strategy is adopted, drawing on the work of Rubin (1987) using five plausible values instead of one for each metric summarised. Rubin (1987) shows that the relative efficiency of an estimate does not improve significantly by increasing the number of imputations used to calculate population statistics (as outlined in section 3.1.6, Appendix 3.2, and Appendix 3.3). Indeed, if the rate of missing information was as high as 50%, Schafer (1999: 7) shows that an estimate based on five imputations has a standard deviation that is only 5% wider than one based on an infinite number of imputations; there is therefore limited benefit in working with any more than five plausible values[3].

Following Rubin's guidelines on multiple imputations, to minimise the error attributed to imputation, every student record in $TIMSS_{2007}$ database has five plausible values for each reporting scale in the survey. To allow analysts to incorporate imputation error into analyses of the achievement data, the recommendation is for each analysis to be replicated five times, using a different plausible value on each occasion. The results should then be combined using Rubin's rules (Rubin, 1987) to summarise the statistics into a single point estimate and standard error that incorporates both sampling and imputation errors. The sampling error is calculated by IEA using Jackknife estimations on the available data to get the *within-imputation* variance; and the imputation error is as calculated by Rubin's *between-imputation* variance. Details on Jackknife estimations follows, with an expansion on Rubin's between imputation variance as presented in Appendix 3.2.

Jackknife estimations

Classical statistical methods rely on basic assumptions concerning knowledge of an underlying distribution and acceptance that observed data come from simple random samples with data points independent of each other and identically distributed (*iid*). When these assumptions are not met fully, or the parametric assumptions are difficult to justify, then analysts can make use of *resampling methods* that preserve whatever underlying distribution is present in the data and allow for estimation of error in the desired statistic. The Jackknife method is one of a number of commonly used approaches that use replication procedures to

[3] In most situations only a small number of imputations are required. A fuller justification for using five imputations (plausible values) is provided in Appendix 3.3

estimate parameters where the assumption of *iid* is violated. The Jackknife method allows estimation of parameters and inference on a wide range of statistics, particularly when confronted with complex survey designs.

The basic principles of the method rely on repeatedly drawing representative samples of the population and computing the desired statistic on each. Each replicate sample provides an estimate of the parameter of interest, and the variability among those estimates gives a measure of the sampling variability. Instead of working with one sample, the technique splits the single sample into multiple, randomly generated sub-samples that are independent and identically distributed in the sample space; the analyst can then deploy traditional methods of analysis on this distribution of parameters.

The general process is optimised when a large number of sub-samples are drawn, generating each sub-sample by removing one unit from the full sample and computing desired statistic – the 'drop one' resampling strategy. For example, let $\hat{\theta}_n$ be an estimator of θ based on n i.i.d. random vectors $X_1, \dots, X_n$, i.e. $\hat{\theta}_n = f_n(X_1, \dots, X_n)$. Let $\hat{\theta}_{n-i} = f_{n-1}(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$ be the corresponding statistic based on all bar the i^{th} observation. If the statistic of interest is the achievement score, then an estimate of the variability within those sub-samples can be derived from the resulting 'simple replicated sample'. Each replicate sample provides an estimate of the mean, and the variability among the replicate samples gives a measure of variance from the overall sample mean, the simple average of the *r* replicate estimators (Rust, 1985: 383). This approach is computationally demanding when *n* is large and there is potential loss of precision when samples are drawn from a complex design, such as the stratified multistage cluster-sampling technique used to select samples of students in TIMSS_{2007}. To accommodate those design concerns, Rust (1985) outlines a commonly used variant on the simple replicated sampling method – a 'random groups method'. The original sample is divided into a number of groups so that each reflects the design of the full sample. An estimate of the population parameter (mean) is formed from each replicate group, and the variability among those estimates is used to get an estimate of the variance of the mean from the full sample, much as outlined above in the 'drop one' resampling strategy. This Jackknife Repeated Replication (JRR) technique is used in TIMSS_{2007} with the samples of schools paired and assigned to a 'sampling zone'. If there is an odd number of schools, once all possible pairs of schools have been identified using known strata and the survey design, students in the remaining school are divided up to make two 'quasi' schools for the purposes of calculating the Jackknife error (standard error of the mean). In the Scottish G4 data there are 70 sampling zones identified; with 65 sampling zones in the comparable G8 data.

In computing the variance, both schools in each sampling zone were randomly assigned an indicator (0 or 1), which determined whether samples of students from those schools were dropped or included when it came to creating the replicate samples. Those that were included had a double weighting assigned to them, i.e. double the original sampling weight for student *i* ($2 \times W_{0i}$); and those students excluded at each wave of replications had their original sampling weight (W_{0i}) recoded to zero. Foy *et al.* (2008) note that in the analyses of $TIMSS_{2007}$, 75 replicate weights were computed for each country regardless of whether or not this exceeded the number of sampling zones in the country. Having a fixed number of replications meant standard errors for several countries could be computed at the same time with no detrimental effect on the magnitude of the error variance; any additional replicate weights, beyond the number of sampling zones in the country, were recoded to the overall sampling weight so that those cases did not add to the variance calculation. To summarise, using the JRR algorithm used in $TIMSS_{2007}$, any statistic *t* from the sample of a country can be computed using the formula for the sampling variance estimate of the statistic *t* is:

$$Var_{jrr}(t) = \sum_{h=1}^{H} [t(J_h) - t(S)]^2$$

where *H* is the total number of sampling zones; taken to be 75 in this survey (as justified above). The term $t(S)$ is the statistic as computed for the whole sample with overall sample weights applied. The term $t(J_h)$ denotes the same statistic using the h^{th} jackknife replicate sample and its corresponding sampling weights, which are identical to the overall sampling weights except for the students in sampling zone *h*. Students from one of the schools in the h^{th} sampling zone are randomly allocated a zero weighting, with students in the other school being allocated a double weighting; all as outlined above. Any calculation in the summation for schools beyond the total number of sampling zones will reduce to zero because $\mathrm{t(J_h)}$ is set to be equal to the statistic as computed for the overall sample weight i.e. $\mathrm{t(J_h) = t(S)}$.

The jackknife estimation was invented by Quenouille in 1949 for the purpose of correcting possible bias in a distribution (small n). In 1958 Tukey noticed that the procedure could be used to construct reasonably reliable confidence intervals for a wide variety of estimators, and so might be viewed as being as useful to a statistician as a regular jackknife would be to an outdoorsman; giving the name to this resampling method (Jiang, 2010). Bootstrap methods were subsequently invented around 1979 by Efron; computationally more intensive (although easier to program) but these methods can give more accurate results in some cases.

Variance estimation can therefore be achieved without detailed knowledge of the sample distribution or involved analysis of the estimation method. We only need a sample and an estimation definition. It cannot be over emphasized that plausible values are not test scores for individuals in the usual sense, and therefore cannot be used to provide feedback, make inferences or decisions about *individual* students. The plausible values provided in $TIMSS_{2007}$ serve only as a means to compute estimates of population characteristics, to permit estimation of population quantities (e.g. regression coefficients), and to provide accurate measures of trends from previous assessments. By using multiple imputations (plausible values), Rubin (1987) and Schafer (1999) show population estimates and conclusions are based on robust data that can provide valid insights to and interpretations of students' and teachers' practices as linked to education achievement data.

3.2.6. Rescaling ordinal polytomous variables

Background data collected from students, teachers and schools is reported in a range of ways. First there are biographical and factual data collected through dichotomous responses or categorical variables. Second there are many background variables in the surveys that seek to measure attitudes or opinions that are based on Likert or ordinal scales. The data associated with those attitudes or opinions are collected as ordered categorical data or are derived from combining clusters into latent variables that have an ordered categorical structure. Decisions have to be made regarding the most suitable metric for those ordinal polytomous explanatory variables, where simple categorical approaches fail to utilise the complete information in the data and there is scope to transform the data to strengthen methods of analysis. Following some introductory remarks on the metric of variables an overview of a rescaling mechanism is provided, building on the work of Snell (1964) to transform the ordinal polytomous structure to follow an underlying continuous scale of measurement. These background data explain variances across the population and are used as explanatory variables in the models of student achievement.

Metric of variables

There have been long-standing debates over the classification of data measurement and the suitability of statistical methods adopted by researchers (Gaito, 1980; Gardner, 1975; Jaeger, 2008; Michell, 1986; Stevens, 1946; Townsend and Ashby, 1984). At the heart of these debates has been the 'permissible statistics' controversy, where the arguments have centred on whether or not the measurement scale and statistical procedures need to be in direct alignment with one another. Stevens' early work presented different types of

measurement scale (nominal, ordinal, interval, and ratio) and clearly associated each type of metric with acceptable statistical procedures (Stevens, 1946). Subsequent debate focused on the ordinal/interval scales and their association with non-parametric/parametric dichotomy. The restriction on permissible statistics rested on the nature of the ordinal measure, where in the strictest sense the calculations associated with means and standard deviations are in error because successive intervals on the scale are unequal in size. Two assumptions fundamental to analysis of variance methods are:

a. The residual deviations are normally distributed, and
b. The residual variances are homogeneous. (Snell, 1964: 593)

When working with data on an ordinal scale, the basic summary statistics of mean and standard deviation ought not to be used, but the distinction between ordinal and interval scales is not always so sharp. Gardner (1975) refers to a 'grey region' lying between the black (scales that approximate well to the requirements of interval strength) and white (rating scales and ranks that are clearly ordinal but not interval) dichotomies of measurement. A major constituent of this grey region are Likert-type scales that feature prominently in the social sciences for measuring attitudes and opinions. Responses to such items reflect the subjective nature of measurement that is based on a respondent's judgment on a scale consisting of several ordered categories.

There are several ways in which Likert scales may be utilised in survey research. First, when a single item is used, Givon and Shapira (1984) suggest three issues need to be considered:

(1) *Whether the attribute being evaluated in one's mind is along a continuous or on a categorical scale*.
(2) *Whether the scale extends infinitely or is bounded.*
(3) *The specific distribution which describes the subject's responses*.

Subjects can find categorical ratings provide a natural and easy form of judgement, but an underlying continuous scale may represent their attitudes more accurately. Cox's review of over 80 years research on the optimal number of response alternatives for a scale (Cox, 1980) notes that the scale needs to be refined enough to be capable of transmitting most of the information available from subjects without being so refined that it simply encourages response error. From the subjects' perspective, error can be induced through being presented with too daunting a range of options or levels of precision. However, it is not solely about maximising the amount of transmitted information because the researcher may not require such precision to address the question under scrutiny. Indeed a crude

measure can be adequate for the researcher's purposes, particularly where: 'the information-processing capacity of the researcher rather than that of respondents may dictate the optimal number of response alternatives' (Cox, 1980: 409). Cox concludes that there is no single number of response alternatives for a scale that is appropriate for all circumstances, but that scales with two or three response alternatives are generally inadequate and those using more than nine provide diminishing returns to the researcher. An odd number of response alternatives are preferable to presenting an even number, permitting a neutral position where that is justifiable; the widely accepted number of seven plus or minus two is recommended to provide subjects with an adequate number of reasonable response alternatives whilst minimising any overuse of the neutral stance. Regardless of the number of response alternatives provided, it is important to distinguish between stimuli where there is a judgement being made that relates to an underlying continuous scale as opposed to a discrete, dichotomous, categorical scale that is more limiting in terms of analysis.

In terms of evaluating attitudes, respondents indicate their rating by placing a mark at an appropriate point on a scale that runs from one extreme of the attribute to the other. This notion of an underlying continuous scale is emphasised when scales are presented through electronic media, with a 'graphic scale' or 'slider control' provided for the respondent to select a value on a scale that is bounded by the concept of 'best' and 'worst' (Page-Bucci, 2003). Similarly when respondents are evaluating a likelihood of an outcome, the probability judgement will be bounded in the [0, 1] range. That said there are advantages in having the scale conceptually unbounded in either direction to permit appropriate analyses and to correct for 'end effects' if the underlying distribution is itself unbounded, e.g. the widely used normal distribution (Cox, 1980:411).

A second way survey research uses these measures is through summated Likert scales. The desired measure may be a result of a subject's response to several items, providing access to constructs that cannot be measured by a single rating scale, or where the researcher chooses to combine responses, by adding the values for each response; hence the name 'summated scales'. This is sometimes referred to as 'score building' (Dittrich *et al.*, 2007) where several items are treated as belonging to the same numerical scale and are either summed over the items, or combined to form a latent variable with a single score produced to measure a common characteristic of that item-cluster for a respondent.

There are several ways such survey data can be managed and organised for subsequent regression analyses. First, the separate categorical responses can be analysed by recoding the categorical variable using traditional dummy coding (0/1) to identify dichotomous variables. In general, a categorical variable with k levels will be transformed

into *k*-1 variables each with two levels; dichotomous variables can then be directly entered into the regression model. When summated Likert scales or derived variables are in operation the number of categories will usually increase. For instance in $TIMSS_{2007}$ the IEA combine responses by reporting the *average* response score across the group of separate items to generate a derived variable for a particular construct; using the average results in non-integer scores e.g. combining four response scores 2,2,3,2 will give an average response score of 2.25. These average values could be analysed using dummy coding as above, but given the number of categories will increase it may become too cumbersome to incorporate a large number of dummy variables, especially for large *k*. An alternative approach, as adopted by IEA in an attempt to simplify the data for wider interpretation and international comparison, is to use an index that assigns respondents to one of three levels of response – high, medium, or low – based on their responses to the component parts. A three-point index is calculated with suitable cut-points from the resultant scale and these three categories can then be analysed using traditional dummy coding to identify dichotomous variables. This is the first method I consider for the explanatory variables in my model of achievement scores, using the same methodology as promoted by IEA for their analysis of TIMSS data.

A second approach is to assume an underlying continuous metric for the construct, based on the extended categorical variable. In this way the derived variable is entered in the regression model as a pseudo-continuous explanatory variable, using the full range of response scores derived from the averages. Third and finally, I consider a rescaling of the original arbitrary scores assigned to the different categories to take account of the observed data and to base the scoring system on the empirical findings. The rescaling mechanism I adopted is as outlined by Snell (1964) who provides a credible approach to rescaling ordered categorical data that is detailed below.

Rescaling mechanism

Arguments to support rescaling of data are presented by Labovitz (1967) where he contends that the advantages of being able to use parametric statistics and more powerful classical techniques outweigh concerns from violating any underpinning assumptions on measurement scales being exactly interval or ratio. His supporting arguments for this position are:

> (1) the insensitivity of ordinal and other nonparametric techniques, e.g., the waste of information by not considering the distance between ranks, (2) the small error that results from assigning numbers to ordinal data and then treating those categories as if they conform to an interval scale, (3) tests of statistical robustness, which have shown that certain tests are interpretable, although selected assumptions are not met, and (4) the power-efficiency of tests. (Labovitz, 1967: 159-160).

Even though the measurement scale of the original data are not exactly interval or ratio, and as such do not satisfy fully the underlying assumptions for parametric statistics, Labovitz argues for the use of parametric tests if the data are *close* to the required form. Labovitz regards it as 'both legitimate and useful' to call upon well-established and clearly interpretable statistical techniques.

A rescaling mechanism can provide the basis of meeting the required assumptions of homoscedasticity and having normally distributed residual deviations. The aim of Snell's paper was to present a scoring procedure that satisfies, as far as possible, these assumptions, so that after the rescaling process one can proceed with classical techniques of analysis of variance. When working with ordered categorical data, traditional approaches use arbitrary scores of 0, 1, 2 … that are subsequently used and analysed on the assumption that necessary conditions for analysis of variance techniques are not violated (Independence, Normality, and Homoscedasticity). These assumptions for ANOVA are broken when dealing with simple integer score responses that can be bunched towards one end of the rating scale, or where the data are otherwise very skew. The theory of Snell's rescaling rests on assumptions of the normal distribution but given the complexity of the mathematical expressions this generates, he suggests a reasonable simplification to base the scoring procedure on a logistic model. The precise procedure requires an iterative solution but for most practical purposes, Snell (1964) contends an approximate solution as adequate, provided there are no obvious irregularities in the data and that estimates based on the weighted sums of the observed data are close to the correct values for the theoretical proportions. Snell's approximate solution, based on observed data, is set out below; I have used this approximate method in all rescaling exercises pursued in my analysis of $TIMSS_{2007}$ data.

The rescaling model assumes there is an underlying continuous scale against which the scale categories represent intervals. Many of the scales used in the survey questionnaires are of a subjective nature, requiring a response or judgment on a scale consisting of several ordered categories. For example:

Agree a lot – Agree a little – Disagree a little – Disagree a lot

At least once a week – Once or twice a month – A few times a year – Never

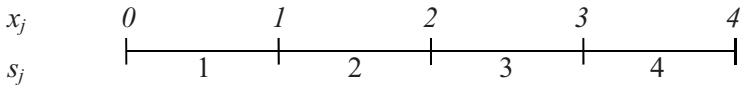
During most lessons – Most weeks – Once or twice each term – Once a year or less

Very confident – Fairly confident – Not very confident – Not at all confident

If a traditional scoring system was adopted, then the scale for four categories would be developed from the following process:

Figure 3.2-6: Intervals defining categories

define points x_j (j= 0, 1, 2, 3, 4) such that category s_j corresponds to the interval x_{j-1} to x_j



To generalise the model there are assumed to be m groups of observations, with the number of categories j extended from 4 to k. The underlying continuous distribution function is denoted $P_i(x_j)$, and the probability of an observation of group i being in category s_j equal to:

$$P_i(x_j) - P_i(x_{j-1}), \qquad i = 1,2, \dots m; \;\; j = 1,2, \dots k.$$

Simplifying the model as suggested above, by basing the procedure on a logistic distribution, the logit of the proportion $P_i(x_j) = a_i + b_i x_j$ and the distribution function probability model is of the form:

$$P_i(x_j) = \frac{e^{a_i + b_i x_j}}{1 + e^{a_i + b_i x_j}}$$

$$P_i(x_j) = \left[\frac{1 + e^{a_i + b_i x_j}}{e^{a_i + b_i x_j}}\right]^{-1}$$

$$P_i(x_j) = \left[1 + e^{-(a_i + b_i x_j)}\right]^{-1}$$

Each observation is characterised by two parameters: a_i , representing the location of the observations, and b_i , their spread. Wishing to satisfy the criterion of homogeneity of variance Snell (1964: 594) takes $b_i = 1$ for all i, giving:

$$P_i(x_j) = \left[1 + e^{-(a_i + x_j)}\right]^{-1} \qquad (1)$$

- a logistic distribution with mean $-a_i$ and variance $\pi^2/3$.

The estimates for parameters x_j and a_i can be found through maximum likelihood methodology. Writing $P_i(x_j)$ as P_{ij} , the logarithm of the likelihood is:

$$L = \sum_{ij} n_{ij} log_e(P_{ij} - P_{i,j-1}), \qquad i = 1, 2, \dots m \; ; \;\; j = 1, 2, \dots k$$

Taking $(P_{ij} - P_{i,j-1})$ term and expanding using (1):

$$P_{ij}\left(1 - \frac{P_{i,j-1}}{P_{ij}}\right) = P_{ij}\left[1 - \frac{1}{1 + e^{-(a_i + x_{j-1})}} . \frac{1 + e^{-(a_i + x_j)}}{1}\right]$$

$$= P_{ij}\left[\frac{1 + e^{-(a_i + x_{j-1})} - [1 + e^{-(a_i + x_j)}]}{1 + e^{-(a_i + x_{j-1})}}\right]$$

$$= P_{ij}\left[\frac{e^{-(a_i+x_{j-1})}}{1+e^{-(a_i+x_{j-1})}}.\left(1-\frac{e^{-(a_i+x_j)}}{e^{-(a_i+x_{j-1})}}\right)\right]$$

$$= P_{ij}Q_{i,j-1}\left[1-e^{-(a_i+x_j)+(a_i+x_{j-1})}\right] = P_{ij}Q_{i,j-1}\left[1-e^{-(x_j-x_{j-1})}\right]$$

Giving

$$L = \sum_{ij} n_{ij} log_e \left\{P_{ij}Q_{i,j-1}[1-e^{-(x_j-x_{j-1})}]\right\} \quad (2)$$

where $Q_{i,j-1} = 1 - P_{i,j-1}$. The logistic distribution has similar properties to the normal curve, matching closely across most of its range and extending in both directions to infinity. Given those end-values we take $x_0 = -\infty$ and $x_k = +\infty$ giving the associated probabilities of $P_{i0} = 0, P_{ik} = 1$. The choice of origin is arbitrary, so I followed Snell's lead by taking $x_1 = 0$ and then used maximum likelihood methodology by differentiating (2) and equating to zero to generate expressions for the remaining estimates x_j $(j = 2, \dots k-1)$ and a_i $(i = 1, \dots m)$. The three expressions used to calculate estimates of the parameters are:

$$0 = \frac{\partial L}{\partial x_j} = \frac{N_j}{\left[e^{(\hat{x}_j-\hat{x}_{j-1})}-1\right]} - \frac{N_{j+1}}{\left[e^{(\hat{x}_{j+1}-\hat{x}_j)}-1\right]} + N_j - \sum_{i=1}^{m}\left(n_{ij}+n_{i,j+1}\right)\hat{P}_{ij}$$

$$\text{for } j = 2, \dots k-2 \quad (3)$$

$$0 = \frac{\partial L}{\partial x_j} = \frac{N_{k-1}}{[e^{(\hat{x}_{k-1}-\hat{x}_{k-2})}-1]} + N_{k-1} - \sum_{i=1}^{m}(n_{i,k-1}+n_{ik})\hat{P}_{i,k-1}$$

$$\text{for j} = \text{k} - 1 \quad (4)$$

and

$$0 = \frac{\partial L}{\partial a_i} = \sum_{j=1}^{k-1} n_{ij} - \sum_{j=1}^{k-1}(n_{ij}+n_{i,j+1})\,\hat{P}_{ij}$$

$$\text{for } i = 1, 2, \dots m \quad (5)$$

To get the approximate solutions, the theoretical proportions $\hat{P}_{ij}$ in (3) and (4) are replaced by the observed proportions p_{ij}. The starting point is to generate a solution for $(x_{k-1} - x_{k-2})$ from (4) and then to use that value in solving (3) to get a sequence of similar expression down to $(x_2 - x_1)$.

Having already defined $x_1 = 0$, all values of x_j $(j = 2, \dots k-1)$ can be found by working backwards through those equations.

Once the class boundaries x_j are estimated, mid-points are taken for the scores (Figure 3.2-6). Class mid-points are easily calculated by averaging the cut-points; extreme categories are obtained using tail values of the distribution function (1). The upper tail $(x\ to\ \infty)$ is equal to:

$$x - (log_e P)/Q$$

where P denotes the probability of a value less than x, and $Q = 1 - P$. The mean value is at a distance $-(log_e P)/Q$ beyond the point x_{k-1}. Similarly for the average value in the lower tail, the mean value will be at a distance $-(log_e P)/Q$ below the point x_1 that has already been defined to be zero. The necessary calculations are completed in an Excel spreadsheet that provides a template for rescaling ordered categorical data from a range of variables. The final step of rescaling adjusts the new scale to centre the data, standardising the scale to have mean zero and unitary standard deviation. This final transformation speeds up the execution of MCMC estimation methods, minimising the number of iterations by having all continuously distributed explanatory variables centred at their mean (Browne, 2012). Having rescaled the variable it is entered in the regression model as a pseudo-continuous explanatory variable.

In summary, there are therefore three ways of handling the ordinal polytomous variables I have encountered in the survey data:

(i) The IEA-style of categorical variables and associated dummy variables based on an index that partitions the variance in the model;

(ii) Considering the extended categorical variable as a continuous derived variable that reflects the conceptual framework and underlying distribution on which respondents base their answer; and finally

(iii) A rescaled pseudo-continuous scale that better reflects the source data as it is based on a scoring system that takes account of the observed data (Snell, 1964).

A number of scenarios are presented in full within Chapter 5: Primary Analysis; the empirical data appears to support use of the third method which is thereafter used to rescale ordinal polytomous variables that may be assumed to have an underlying continuous metric.

3.3. Methodological issues related to analyses of survey data

Methodological issues pertaining to the analysis of hierarchical clustered data focus on the choice of dependent variable and the unit of analysis in any model of students' achievement that is developed. An outline of the primary method of analysis adopted in this study is presented before providing an overview of three alternatives that are subsequently evaluated on empirical grounds. The primary approach takes on a new multilevel plausible value method that exploits the hierarchical structure of the data sets by building a multilevel model of student's achievement with 'plausible value' as the unit of analysis. This extends

the structure of the hierarchical model to have four levels, with 'plausible value' nested within 'student', which in turn is nested within 'teacher' and 'school' levels. Three alternative options discussed are first, to use an arithmetic average of the five plausible values as the dependent variable in models of student's achievement; second, to use a single plausible value as the dependent variable in the analysis of student's achievement; and third, to separately model all five plausible values before merging the resultant statistics as recommended by Rubin (1987) to combine imputed data. The findings from the primary method of analysis, using a multilevel model with plausible value as the unit of analysis, can be compared with the findings from each of the three alternative methods to determine whether or not substantive conclusions differ; comparison with Rubin's method is of primary interest. Recommendations will be made over use of one approach in preference to others.

Throughout the analysis, substantive conclusions are reported in terms of statistical significance of the contributing variable, as well as assessing the practical significance based on the raw difference identified through the regression coefficient. Assessing the relative size of effects is pursued through use of standardised estimates of effect size. An overview of these approaches to assessing the strength of association with achievement scores in the multilevel models will conclude this section.

3.3.1. Multilevel analyses

In order to confirm the data are suitable for a multilevel analysis, a basic model was developed on the basis of the hierarchical structure discussed in section 3.2.1. First the data were transformed to create a long file with plausible value as the unit of analysis. A general model of student achievement was set up as an unconditional model, and then predictor variables as related to literature review and EDA were considered for inclusion, adding and retaining if providing a significant explanation of variance in the overall model of achievement. This section outlines those steps in a general capacity with the resultant models reported within the primary analyses of Chapter 5.

A number of simple models are used to discuss the principles and adopted practices of modelling the data with different levels of classification. Clearly the simplest situation is the null model with one level of variance – describing the data in terms of an overall mean β_0 plus a student residual error (e_i) that is itself normally distributed with $e_i \sim N(0, \sigma_i^2)$. This gives a very basic model with all the variance attributed to the student (i); a fixed effects model:

$$\text{ASMMAT\_AV}_i = 497.119(1.201) + e_i$$

$$e_i \sim N(0, \sigma_e^2) \quad \sigma_e^2 = 5635.257(127.548)$$
$$-2*loglikelihood = 44797.131(3904 \text{ of } 3904 \text{ cases in use})$$

This structure can readily lead to increasingly more sophisticated models of achievement where there are two nested levels, taking individual students within schools say, allowing the researcher to comment on differences between schools as well as differences within schools (between students). The structure of the equation now takes the form:

$$ASMMAT\_AV_{ij} = \beta_0 + u_{0j} + e_{ij}$$

As before, this represents an overall mean (β_0) but the equation now has two variance components, one for the differences between schools (u_{0j}) and the other for differences within schools (e_{ij}) which represents the between student differences. The model of student achievement is calculated using MLwiN to give the random intercept model for school variation:

$$\text{ASMMAT\_AV}_{ij} = \beta_{0j} + e_{ij}$$
$$\beta_{0j} = 493.604(2.795) + u_{0j}$$

$$u_{0j} \sim N(0, \sigma_{u0}^2) \quad \sigma_{u0}^2 = 853.961(128.445)$$
$$e_{ij} \sim N(0, \sigma_e^2) \quad \sigma_e^2 = 4770.469(109.891)$$
$$-2*loglikelihood = 44380.941(3904 \text{ of } 3904 \text{ cases in use})$$

Having included the school level of analysis, the model is a two-level variance components model with the overall mean of the dependent variable ASMMAT_AV defined by a fixed coefficient β_0 plus two variance components that gives the model its name. The original student variance has been split into a school variance and a reduced individual contribution as described above.

The second level of the model allows the mean of the j^{th} school to be raised or lowered by u_{0j} that is Normally distributed with a mean of 0, and an estimated variance of 853.96. Convention dictates that *u* is used to denote the 'level 2' residual, which in this case is the school effect on the model. The variance partition coefficient (VPC) is calculated as:

$$\frac{853.961}{853.961 + 4770.469} \cong 0.152$$

This shows that 15% variation in the model is attributed to between schools.

Suppose $ASMMAT\_AV_{ij}$ is the observed value for the i^{th} student within the j^{th}

school and that $\widehat{ASMMAT\_AV_{ij}}$ is the predicted value from the regression – in this case it will be the overall mean for ASMMAT_AV plus a variance component for the school. Typical values can be confirmed in the data view of the file in Figure 3.3-1:

Figure 3.3-1: Predicted values from regression

Data

goto line 1 | view | Help | Font | Show value labels

	IDSCHOOL(138)	ASMMAT_AV(138)	Pred_Null(138)
1	1.000	496.429	539.915
2	2.000	498.258	510.860
3	4.000	397.048	505.997
4	5.000	426.384	479.835
5	6.000	459.315	500.360
6	7.000	600.571	521.039
7	8.000	540.547	548.552
8	9.000	387.026	542.845
9	12.000	547.968	486.449
10	14.000	418.024	479.437

The *raw residual* for an individual is the difference between observed and predicted score (ASMMAT_AV) i.e. $r_{ij} = y_{ij} - \hat{y}_{ij}$, where the *raw residual* for the j^{th} school is written as r_{+j}, the mean of r_{ij} for the students in that school. The estimated level 2 residual for the school is obtained by multiplying r_{+j} by a shrinkage factor, a transformation that gives a *shrunken residual.* The shrinkage factor *k* is $\frac{\sigma_u^2}{\sigma_u^2+(\sigma_e^2/n_j)}$, where n_j is the number of students in school *j*, giving:

$$\hat{u}_{0j} = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma_\epsilon^2/n_j} r_{+j}$$

The within-student distribution stays the same (σ_e^2 is constant) but if more students are sampled ($n_j$increases), then σ_e^2/n_j will get smaller and so *k* will get closer to 1. In other words, there will be less shrinkage if more students are sampled. The shrinkage factor takes account of the number of cases within that level and correspondingly the residuals are affected by the size of each group. This means that, for level 2 units or higher, where the information is scarce, the estimate of the regression line for that unit is 'shrunk' towards the mean regression. For level 2 units with larger sample sizes, the estimates of the regression coefficients are more reliable and level 2 residuals are 'less shrunk' towards the mean.

MLwiN calculates the residuals at each level, taking the necessary shrinkage factor into account, and provides options for a variety of graphical displays to inspect them. To check the normality assumption, a normal probability plot of ranked residuals against a normal curve can be inspected. If normality assumption is valid then the points should lie

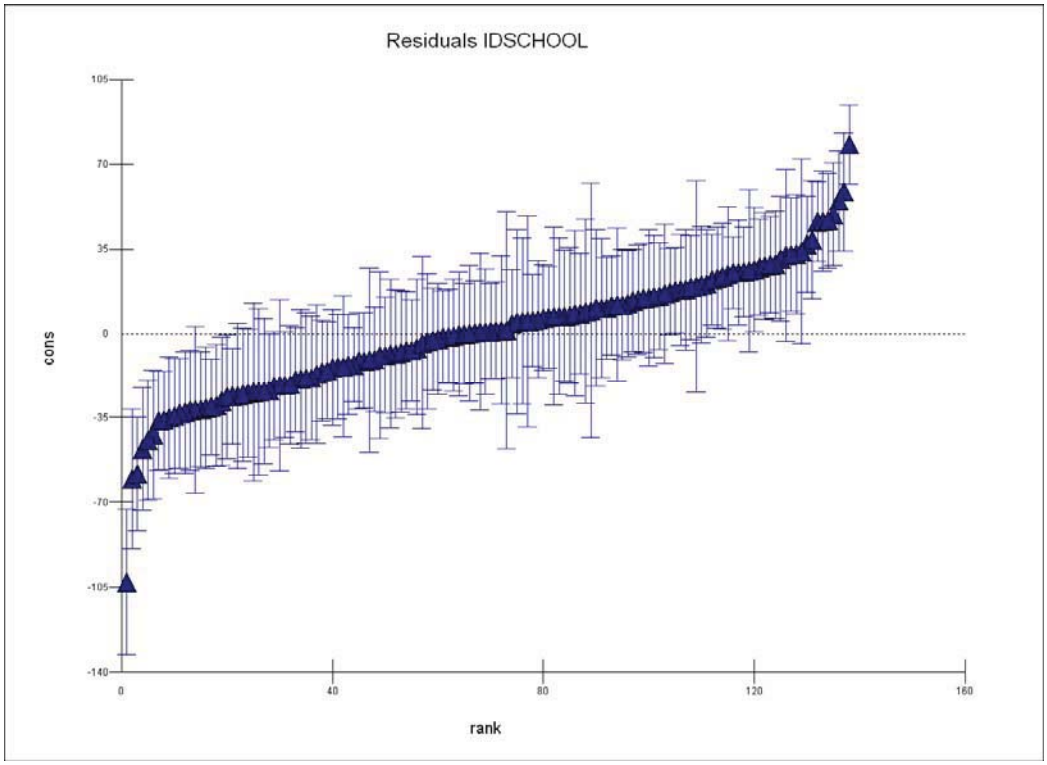
approximately on a straight line. Plotting the residual $+/-1.96$ sd by rank produces a caterpillar plot, as illustrated for this model in Figure 3.3-2. If the plots do not overlap zero, as witnessed at either ends of the plot for the low and high ranks, those residuals represent school departures that are significantly different (at the 5% level) from the overall average predicted by the fixed parameter β_0.

To test the significance of the school effects a likelihood ratio test can be carried out. This test takes the difference between the *-2*log likelihood* values for the two models and compares the result with a chi-squared distribution on 1 degree of freedom, given there is one additional parameter in the second model (σ^2_{u0}). The significance of the school effect is:

$$LR\ test = 44797.131 - 44380.941 = 416.19$$

A chi-squared distribution on 1 degree of freedom is 3.84 so there is clear evidence of school effects on achievement.

Figure 3.3-2: Caterpillar Plot for Residuals (IDSCHOOL)



For comparison purposes, when teacher (IDTEACH) is used as the second level of the hierarchical structure, the new parameters in the model are:

$$ASMMAT\_AV_{ij} = \beta_{0j} + e_{ij}$$
$$\beta_{0j} = 496.020(2.532) + u_{0j}$$

$$u_{0j} \sim N(0, \sigma^2_{u0}) \quad \sigma^2_{u0} = 1277.159(143.987)$$
$$e_{ij} \sim N(0, \sigma^2_e) \quad \sigma^2_e = 4445.557(104.030)$$
$$-2*loglikelihood = 44282.211(3904 \text{ of } 3904 \text{ cases in use})$$

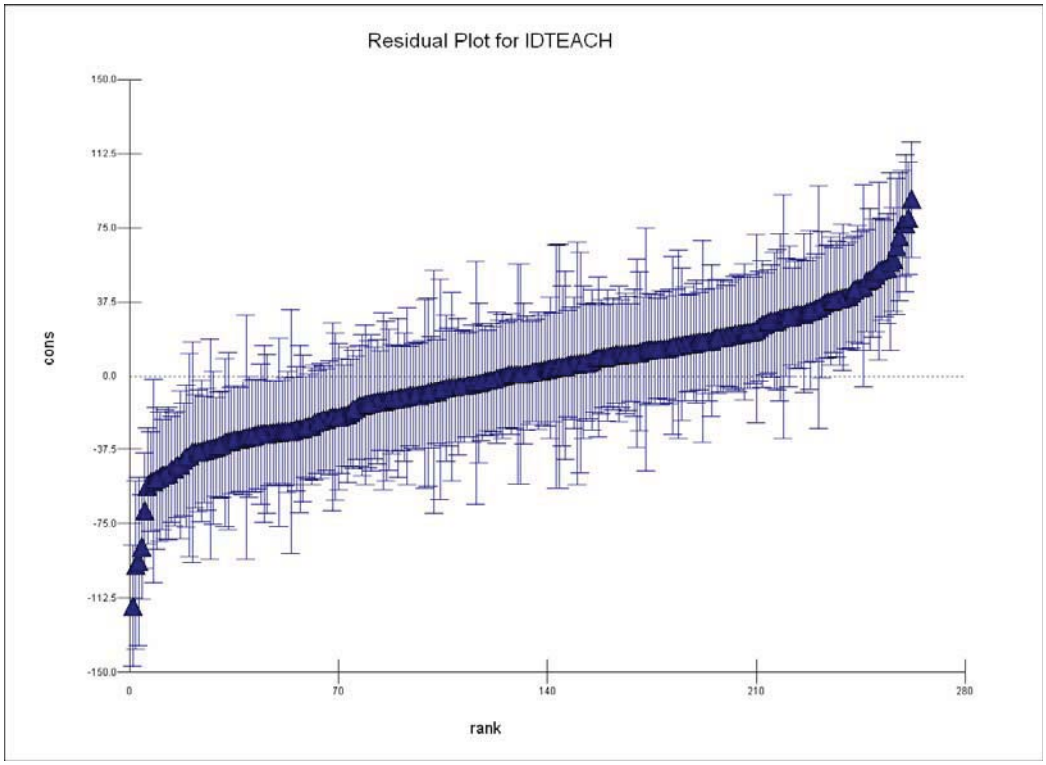
Direct comparison with the school effects model shows that a much greater variation is attributed to the teacher component (σ^2_{u0}), but this needs to be interpreted carefully as the residuals will be affected by the size of each group – the residual estimates can vary greatly in their width because of small samples of students working with some teachers.

The residual plot for IDTEACH (Figure 3.3-3) has similar characteristics to that for IDSCHOOL. The discernible deviations from the mean at either end of the ranked scores can be seen to be significantly different given the 95% confidence intervals do not overlap with zero. Carrying out a similar analysis as above the VPC for teacher-level of the model is:

$$\frac{1277.159}{1277.159 + 4445.557} \cong 0.223$$

This shows that 22% variation in the model is attributed to difference between teachers.

Figure 3.3-3: Caterpillar Plot for Residuals (IDTEACH)



As before, the significance of the teacher effects can be considered using the likelihood ratio test. The significance of the teacher effect in the model is:

$$LR\ test = 44797.131 - 44282.211 = 514.92$$

This test statistic confirms there is overwhelming evidence of teacher effects on achievement. The subsequent analyses will strive to identify some specific aspects of teacher and school effects that can offer explanation for those significant variances.

3.3.2. Primary method of analysis

The primary method of analysis draws upon Bayesian principles and techniques. The survey data are complex in structure and so too are the multilevel models of student achievement that attempt to describe the hierarchical data and to explain variances across a range of parameters. A primary consideration in favouring Bayesian inference over frequentist inference rests on the fact the distributions of selected explanatory and control parameters are unknown, leaving estimates to be determined through observed data only:

> The Bayesian approach delivers the answer to the right questions in the sense that Bayesian inference provides answers conditional on the observed data and not based on the distribution of estimators or test statistics over imaginary samples not observed. (Rossi *et al.,* 2005:4)

Hall (2012) outlines a number of advantages of Bayesian inference over frequentist inference, many of which support the current problem that investigates relationships within hierarchical data. One particular advantage he cites concerns comparison of models, where the goal is to identify an appropriate model of the achievement data. Model comparison using Maximum Likelihood (ML) Iterative Generalised Least Squares (IGLS) methodology is fairly straightforward when comparing simple nested models. Assessment of candidate models can be carried out as a sequence of comparisons between pairs of models. Taking two such models denoted M_1 and M_2, with p_1 and p_2 parameters respectively, the ML estimates under each model can be compared and tested for significant difference. The comparison is similar to a classical testing problem with the null hypothesis of 'no difference' between likelihoods, leading to the likelihood ratio test statistic $2[L(\hat{\theta}_2) - L(\hat{\theta}_1)]$, which for large samples will follow a χ^2distribution with $p_2 - p_1$ degrees of freedom. The change of deviance as measured by the log-likelihood statistic is used as a measure of fit, with any significant reduction in deviance favouring model M_2.

Kuha (2004) challenges the appropriateness of standard tests for all models, particularly where: the models are not straightforward; analysis is extended over large sample sizes; or where comparisons are not restricted to nested models. Significance tests on large samples are sensitive to small deviations from the null hypothesis, a feature that

renders all reasonably parsimonious models as having a statistically significant lack of fit and subsequent rejection from consideration. Kuha also cites standard tests as mostly unsuitable for comparing non-nested models, and that they 'provide little guidance for choosing between models that have not been rejected' (Kuha, 2004: 188). These limitations led researchers to consider other approaches to model comparison and selection, with a range of *penalized* model selection criteria that take sample size and complexity of the model into consideration, trading these against the fit determined by the likelihood ratio test statistic. Two well-used members of this class of penalized models are the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). These are defined as:

$$AIC = 2\left[L\left(\hat{\theta}_2\right) - L\left(\hat{\theta}_1\right)\right] - 2(p_2 - p_1)$$

and

$$BIC = 2\left[L\left(\hat{\theta}_2\right) - L\left(\hat{\theta}_1\right)\right] - \ln(n)(p_2 - p_1)$$

In each case the first term reflects the 'fit' of the two models to the observed data (i.e. the difference of the maximised log-likelihoods); a measure that favours larger models. The second term is the penalizing component, using the number of parameters as a measure of 'complexity' that reduces the information criterion by a differing factor for each of AIC and BIC (i.e. $p_2 - p_1$ is the increased complexity of M_2 over M_1). The two terms pull in opposite directions, with a trade-off between 'fit' and 'complexity'. The comparison favours a parsimonious model unless the 'fit' achieved by a more complex model is sufficiently large. The theoretical differences between AIC and BIC are discussed by Kuha (2004), Spiegelhalter *et al.* (2002) and Hall (2012), with an emphasis on the different assumptions and goals that underpin those statistics. In summary, AIC and BIC have the same aim and common goal to identify a good model; the differences lie in their definition of a 'good model'. The motivation behind BIC is to identify a model with the highest probability of being the true model for the data, working on the assumption that one of the models under consideration is 'true'. On the other hand, the derivation of AIC explicitly denies the existence of any identifiable 'true' model, favouring instead to rely on predictive power of future data as a measure of adequacy for a model.

Bayesian inference allows comparison of models using deviance statistics. Alston *et al.* (2005) suggest a natural approach to compare models is on the basis of their posterior probability distributions. The choice between M_1 and M_2 can be made on the basis of the ratio:

$$\frac{P(M_2|y)}{P(M_1|y)} = \frac{P(M_2)}{P(M_1)} \times \frac{P(y|M_2)}{P(y|M_1)}$$

where larger values would support M_2 over M_1. The second term in this expression,

the ratio of the marginal likelihoods, denoted B_{21}, is called the *Bayes factor* (Kass and Raftery, 1995). In words we have:

$$\text{posterior odds } = \text{prior odds} \times \text{Bayes factor}$$

Unlike the likelihood ratio, instead of maximising, the terms of the *Bayes factor* are obtained by integrating over θ, such that:

$$P(y|M_i) = \int P(y|M_i, \theta_i).P(\theta_i|M_i)d\theta_i \qquad i = 1,2$$

Because a Bayes factor is often difficult to evaluate, a popular alternative is to adopt an approximation to the Bayes factor using one of a number of Information Criterion (IC) statistics. AIC and BIC have been outlined above; Spiegelhalter *et al.* (2002) introduce another variant on this statistic for model comparison, called the Deviance Information Criterion (DIC). Their paper suggests Bayesian measures of complexity and fit can be combined to compare complex hierarchical models, building on the classical estimation of 'fit' and 'complexity' as featured in AIC (and BIC). DIC is analogous to AIC in that it is not based on any assumption of a 'true' model and is primarily concerned with short-term predictive ability. The DIC comprises terms that are a function of the data alone and a measure of the complexity of the model.

Spiegelhalter *et al.* (2002) propose a Bayesian model comparison based on:

$$DIC = 'goodness of\ fit' + 'complexity'$$

$$DIC = D(\bar{\theta}) + 2pD$$

Where $D(\bar{\theta}) = -2\log L(data|\theta)$ and $\text{pD} = \overline{\text{D}} - \text{D}(\bar{\theta})$ i.e. the 'posterior mean deviance' minus 'deviance calculated at the posterior mean of the parameters'. This makes pD a measure of complexity that estimates the 'effective number of parameters'. Models with smaller DIC are better supported by the data.

As shown in Spiegelhalter *et al.* (2002), the DIC is a generalisation of the well-known AIC. Taking the definition of $AIC = D(\hat{\theta}) + 2p$, where $\hat{\theta}$ is the maximum likelihood estimate of the parameter vector and p the number of parameters, the similarities can be highlighted. For non-hierarchical models, $p \approx pD, \hat{\theta} \approx \bar{\theta}$ and therefore $AIC \approx DIC$.

Hall presents similar arguments against use of standard significance tests, claiming them as inappropriate for complex models, non-nested data or hierarchical data, stating: 'frequentist model fit statistics cannot compare different methods or hierarchical models' (Hall, 2012:22). This argument supports development and use of Bayesian techniques. Hall further argues in favour of Bayesian inference via MCMC, an approach that allows complicated models to be estimated, going well beyond anything that frequentist inference could accommodate. If the chained run is long enough, there is a theoretic guarantee that the

Bayesian inference will converge, whereas frequentist approach has no guarantee of convergence. One downside of Bayesian inference is the long run-time required to reach a state of equilibrium that offers confidence and accuracy sought for the model of interest; in complex models with large samples as witnessed within $TIMSS_{2007}$ data sets can take many minutes to converge given the high number of iterations required to satisfy the diagnostic analyses (~250,000+).

Kuha (2004) warns against over-reliance on any one information criterion statistic, and recommends multiple analyses that draw on different IC statistics as well as other indices and models of fit before reaching any conclusions on model comparison. When a range of criteria agree on an optimal model then that clearly further supports conclusions and offers robust assurance on decisions. Where there is disagreement, analysts must take the underlying conditions, assumptions and purposes into account, for instance what is the nature of 'true' models in the field of study? The three IC statistics outlined above can answer different prediction problems. The BUGS (Bayesian inference Using Gibbs Sampling) project outlines a very comparable scenario to my data by way of explaining which IC is best suited to a problem. Adapting their example to reflect the TIMSS data set, gives:

> Suppose the data are nested on three levels with *students*, nested within *classes*, within *schools*. Then
> (1) if we were interested in predicting results of future *students* in those actual classes, then DIC is appropriate (i.e. the random effect themselves are of interest);
> (2) if we were interested in predicting results of future *classes* in that school, then marginal–likelihood methods such as AIC are appropriate (i.e. the population parameters are of interest);
> (3) if we were interested in predicting results for a new *school,* then BIC/ Bayes factors are appropriate (i.e. the 'true' underlying model is of interest).

In practice, the EDA of survey items identified control and explanatory variables for consideration in models of student achievement. These variables and derived variables are explored in multilevel analyses to confirm suitability for inclusion in subsequent models of mathematics and science achievement. A range of individual student-, teacher- and school-level control variables are considered and retained as warranted, checking variance partition coefficients, significance of coefficients for parameters (significance determined on value of t-test statistic, where t-value>2.58 is significant at 1% level; t-value>1.96 is sig. at 5% level), and overall model comparisons assessed using DIC as dictated by Bayesian methodology and estimates generated through MCMC. Each stage of development is based on initial estimates for predictor and control variables using full iterative generalised least squares (IGLS) methodology, followed by MCMC estimation using extended chains (250,000+ iterations) to satisfy accuracy statistics presented in MCMC diagnostic analyses.

3.3.3. Alternative methods of analyses (empirical perspective on theory)

Three possible alternative analyses will be reviewed to determine whether they generate the same conclusions as drawn through the primary Bayesian approach. The first alternative is to empirically check findings generated through using the average of the five plausible values as the dependent variable. Such an approach has been criticised on theoretical arguments, where an explicit variance component associated with imputation is not included. This study provides an opportunity to compare outcomes and weight of evidence on empirical grounds where the variance components are accounted for at each level of the hierarchical model, including variance within plausible values assigned to each student. As a second alternative I will evaluate the more extreme case of developing multilevel models of achievement based on a single plausible value. This will again provide empirical evidence of practice and will address the issues of whether any valid and significant conclusions can be drawn from using a narrow subset of the achievement data. Third and finally, I will compare earlier findings with the method recommended by IEA and other analysts who choose to follow Rubin's recommendations for handling analyses based on multiple imputation and plausible value methodology; namely evaluating five separate models, one for each plausible value, and thereafter combining the results. The direction and strength of effect as determined through Rubin's method is compared with that generated through the other methods under consideration, inspecting observed changes in coefficient estimates, standard errors and effect sizes, noting instances where conclusions on substantive issues need to be revised in light of the evidence provided. Recommendations on suggested practice will emerge, taking account of precision of findings and level of workload required to reach conclusions that satisfy substantive concerns.

3.3.4. Effect size

The use of effect sizes in reporting research has not as yet been universally accepted with much debate centred on the various forms, presentations and appropriateness of this statistic in published research (Coe, 2002; Lipsey *et al.,* 2012; Schagen and Elliot, 2004; Tymms *et al.*, 1997). A driving force behind increased use of effect size is to improve communication and interpretation of research findings, so that data and findings are presented in a useful and understandable format for policy and practice communities. The American Psychological Association (APA) 'encouraged' effect size reporting in its *Publication Manual* (fourth edition, 2004). In the sixth edition (Fidler, 2010), the association

more forcibly spelled out the merits of reporting effect size with appropriate confidence intervals in an effort to move away from Null Hypothesis Significance Testing (NHST) that has dominated 'over six decades [of] psychology and many other life and social sciences' publications. The emphasis on providing a measure of statistical uncertainty has also been made by Coe (2002) and several contributors to the volume titled '*But what does it mean*? *The use of effect sizes in educational research*' (Schagen and Elliot, 2004).

A standardised mean difference effect size statistic is commonly referred to as the 'effect size' in a model. This statistic is a simple way of quantifying the difference between groups and as such can serve as a measure of the effectiveness of a factor in any of the multilevel models that I develop. Following Coe (2002) the standardised effect size is basically the difference between mean values, divided by the pooled standard deviation. The literature sets out this calculation in terms of experimental and control groups, but the equivalent in multilevel modelling will be a measure of difference in means between randomly assigned memberships of sub-groups. If a full population is chosen and if there is a random assignment of say 'working in groups', then the effect size is calculated using the population standard deviation, or more strictly the pooled standard deviation of the randomly assigned sub-groups. However if a particular sub-group was identified using background characteristics, and the same random assignment was analysed, then the pooled standard deviation for that restricted group would be used to calculate the effect size. This would be a smaller standard deviation than used above for the full population leading to a correspondingly bigger effect size as a direct result of working with a restricted group. In order to make sensible comparisons between models and to ground the effect size calculations for meta-analyses, it would make sense to use the variance from the null model rather than any model that includes control variables.

The notion of reporting effect sizes is not new, but continued arguments centre on which type of index to use and whether variants on a pooled standard deviation are appropriate. In relation to multilevel modelling as used in this study, following Sammons *et al*. (2002) and Tymms (1997, 2004), the fixed part of the model is essentially the mean difference between two groups after controlling for all other factors. With dichotomous and categorical (dummy) variables the effect size is therefore the regression coefficient divided by the pooled standard deviation. Strictly this means the divisor should pool the raw standard deviation for the intervention and control samples, even when dealing with complex structures in multilevel models; the purpose being to get the best estimate of the respective population standard deviation by using all the data available. However in multilevel modelling the variance for the population of students that is relevant for standardising the

effect is partitioned into between and within components: in this study I have between school variation, between teacher variation, and within teacher variation that is reported as between student variation. In my primary analysis I also have within student variation that accounts for variation between plausible values, but it is the population of students that is reported upon so our interest stops at that level. This makes the between student variation the best estimate of the population variance and standard deviation. For this reason the conventional effect size would standardise effects on the basis of the pooled student-level standard deviations, i.e. the square root of the within group variance in the null model (as argued above). This is simply σ_e, making the effect size $\Delta = \frac{\beta_i}{\sigma_e}$ where β_i is the coefficient associated with the intervention (Tymms *et al.*, 1997).

When a continuous variable is considered in the model, the approach taken follows the principles presented by Glass *et al.* (1981), where they suggest:

$$\Delta = \frac{2zr_{xy}}{\sqrt{(1 - r_{xy}^2)}}$$

Where: r is the correlation between variables x and y;

z is the unit normal deviate at the p^{th} percentile.

An effect size is found by considering it as though it was a dichotomous variable and deciding where to slice the continuous variable to form a discrete equivalent ('discretise' it). Fitz-Gibbon and Morris (1987) simplify the above by choosing the cut points to be one standard deviation above and below the mean, giving:

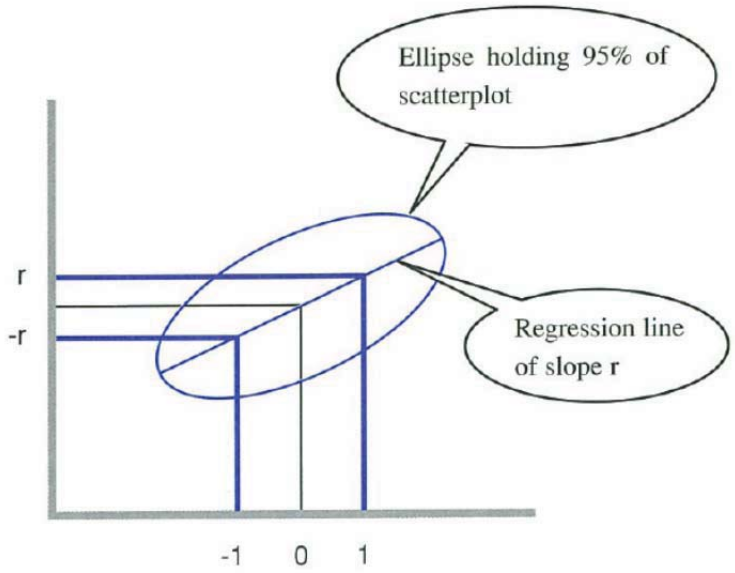
$$\Delta = \frac{2r}{\sqrt{(1 - r^2)}}$$

Tymms (2004: 62) provides a helpful graphic in support of this expression for the effect size (Figure 3.3-4). The scatterplot is of two normally distributed variables each with a mean of 0 and SD of 1. The slope of the line is equal to the correlation coefficient (r). Vertical lines have been drawn to indicate the points that are one SD above and below the mean of 0, tracking across to the values of $\pm r$ given the gradient as r.

Where the multilevel model has a continuous predictor and outcome that follow a normal distribution, or have been normalised (mean=0; SD=1), the coefficient is equivalent to r, making the formula for effect size:

$$\Delta = \frac{2\beta_i}{\sigma_e}$$

Figure 3.3-4: Graphical representation of effect size (Tymms, 2004)



Tymms notes that if the predictor and criterion are not normalised the slope of the line is β_i and the vertical lines in Figure 3.3-4 will meet the horizontal axis at one $SD\ (predictor)$on either side of the mean. The distance between the tracked lines that cut the vertical axis will now be $2\beta_i \times SD\ (predictor)$, giving the more complex formula for effect size as:

$$\Delta = \frac{2\beta_i \times SD(predictor)}{\sigma_e}$$

Any continuous predictors in my models have been normalised to avoid this additional complexity.

When considering the margin for error in estimating effect sizes, Coe (2002:8) argues for a confidence interval since it keeps the emphasis on the effect size, rather than a p-value. He goes on to cite a formula for calculating the confidence interval for an effect size *d* as given by Hedges and Olkin (1985: 86) as:

$$\sigma[d] = \sqrt{\frac{N_E + N_C}{N_E \times N_C} + \frac{d^2}{2(N_E + N_C)}}$$

Where $N_E and N_C$ are the numbers in the experimental and control groups respectively. The 95% CI for *d* would then be calculated from:

$$d - 1.96 \times \sigma[d] \qquad \text{to} \qquad d + 1.96 \times \sigma[d].$$

An alternative approach presented by Tymms (2004: 58) focuses on the *error* of the effect size as calculated by combining the errors from both the coefficient and the SD. If the

error in X is $errX$ and $X = \frac{A}{B}$ (or $X = A \times B$) then in any error analysis:

$$\frac{errX}{X} = \sqrt{\left(\frac{errA}{A}\right)^2 + \left(\frac{errB}{B}\right)^2}$$

This can be computed quite readily from the multilevel output data to give errors of:

$$errX = X \times \sqrt{\left(\frac{errA}{A}\right)^2 + \left(\frac{errB}{B}\right)^2}$$

This formula provides a suitable measure of uncertainty for the effect size and is the format used to report error bounds on computed effect sizes in my multilevel models of achievement. Illustrative approaches for those calculations are presented in Appendix 3.6.

The inclusion of effect sizes with accompanying measures of uncertainty do not replace other well-recognised statistics used in reporting findings but rather complement assessments of significance, publication of raw mean differences/ regression coefficients in multilevel models, and measures of percentage variance explained. These statistics are included in outputs, noting the importance of providing effect size even where significance level has not been met. The p-values on their own can be misleading, especially where there are large samples involved where almost any difference will be statistically significant. Using effect sizes to complement interpretation of findings is essential where comparing the relative contribution of coefficients from different variables that do not report on the same scales. Having a normalised, unit-free measure satisfies the challenge of evaluating effect of different variables by making the outcomes directly comparable across the whole population, placing an emphasis on quantifying an effect rather than merely reporting statistical significance.

Cohen (1969) sought to interpret such quantification within psychological research, by tabling bands of effect size as small, medium and large, with cut points of 0.2, 0.5 and 0.8 respectively. This was not intended as a rigid formulation of effect but in many quarters this classification has been accepted as a 'standard' interpretation. Coe (2004) rightly criticises this stance as being overly simplistic and failing to take account of wider considerations such as cost benefits and practical feasibility in relation to other effects produced by comparable interventions. In the same volume, Goldstein (2004) urges users to consider mechanisms that support interpretation of effects, going beyond a single number summary by assuming responsibility for deciding how to make comparisons that take account of social utility and resource cost. This will involve placing relative costs on competing effects – with analysts then making professional judgements over cost benefits in tandem with reported effect sizes.

To assist with such judgements a range of related scales are suggested by Coe

(2004), linking effect size to known contexts including education grade equivalent (GCSE), educational performance in tests as typical for a year of schooling, and others that relate to probabilities and percentiles as computed from a standard Normal distribution. Effect sizes provide a direct measure of difference as measured in standard deviations. For instance an effect size of 0.35 means the score of the average person in the experimental group is 0.35 standard deviations above the person in the control group. But effect size is equivalent to a z-score of a standard Normal distribution so that same effect size of 0.35 means the average person in the experimental group will exceed 64% of those in the control group i.e. $P(Z < 0.35) = 0.64$. It should be noted however that interpretation of effect size in terms of percentiles assumes normality and is very sensitive to violations of that assumption. The diagnostic checks on variables within the $TIMSS_{2007}$ data support normality, so as a means of interpreting effect sizes I will offer that parallel interpretation of the percentage of control group who would be below the average person in experimental group.

4. Exploratory Data Analysis of $TIMSS_{2007}$ Questionnaires

In addition to the achievement data on Mathematics and Science topics reported in the $TIMSS_{2007}$ survey, the participating students provide background data on their learning context as based on home and school experiences including background attitudinal data about their learning of mathematics and science and academic self-concept. Teachers and schools submit similar data about their teaching of mathematics and science, addressing both curricular coverage and pedagogies. School personnel provided information on school staffing and resources, as well as individual and institutional contextual data that include facts on teachers' professional training and education and views on school ethos and culture. Those data are gathered through questionnaires completed by students, their teachers (G4 general and G8 specialist) and their schools (through administrator or Head Teacher response). The latter data from teachers and schools are collected and reported as a 'student attribute', directly linked to the student's response. Although school-level analyses where the schools are the units of analysis can be performed, the IEA recommend analyses of school-level variables as attributes of students. For example analyses with teacher background data seek to make statements *about students whose teachers have a given characteristic*, rather than about teachers with a given characteristic. E.g. investigating the percentage of eighth-grade students according to the age of their mathematics teachers:

Table 4.1-1: Percentage of Students by Teacher Characteristics (S.E.)

Gender				Age							
Female		Male		29 Years or Under		30–39 Years		40–49 Years		50 Years or Older	
58	(3.1)	42	(3.1)	16	(2.1)	25	(3.0)	25	(2.9)	33	(3.6)

The background data provide a rich source of information for the analysis of learners' experiences in relation to $TIMSS_{2007}$ achievement scores and recent / current policy initiatives within mathematics and science education.

Exploratory analyses of the data, to ascertain distributions within responses for each variable as well as investigating any association with achievement data are summarised in this chapter with supporting data presented in Appendix 4. First the student responses to the background questionnaire are explored. These data reflect a mix of fact and opinion related to personal circumstances at home and outside school, and their experiences of mathematics and science in school. Second the data collected from teachers of mathematics and science who teach the sampled students are described and analysed. Information pertaining to

teachers' personal and professional background is considered in addition to the teachers' responses to themes on instructional practices and attitudes toward teaching mathematics and science. Given those data are gathered as another 'student attribute', it is permissible to investigate relationships with student level data and to consider associations with student achievement. Third and finally for this exploratory data analysis (EDA), information about the student's school is considered. As above, those data are reported as 'student attributes' and will be analysed on that basis, with the responses from their school principal providing both factual information and professional opinion on school characteristics and the national education system experienced by the student. The EDA on school characteristics and students' experiences in those settings will flag up potential associations with student achievement that can be considered further in later analyses. Such analyses will be indicative of associations in the data, offering insights to which features will be worthy of closer inspection and more accurate scrutiny through subsequent modelling of the data that will take full account of the clustered data structure. An arithmetic average of the five plausible values for mathematics, science or their relevant sub-scales is used as a measure of location for student achievement when investigating association between achievement scores and the background variables of interest.

The following sections will also illustrate details of methods used to compute derived variables, reduce data dimension through Principal Components Analysis (PCA), and rescale data using Snell's scaling procedure for ordered categorical data (Snell, 1964; outlined in Chapter 3.2.6). These processes will be outlined and illustrated in full on the first opportunity that arises, thereafter presenting finalised format without a detailed breakdown of intermediate steps for every variable or situation.

4.1. Student Questionnaire (G4)

Descriptive statistics are presented using weighted data, with total student weights (TOTWGT) being applied before computing analyses or drawing graphs of relationships between variables. The rationale behind weighting of survey data is discussed more fully in Chapter 3. Data presented in tabular format have been rounded to 2 effective digits (Chapman, 1996; Ehrenberg, 1986); a feature that makes it easier to read the demonstration tables but one that may result in components not adding to totals because they have been rounded independently. The broad themes of relevance and interest to analysis of Scottish data and its comparator nations are presented under headings of biographical data (gender, home environment, resources and support for students), out-of-school interests, classroom experiences, ICT experiences, and school culture as perceived by the student cohort.

4.1.1. Gender

An even gender balance is reported with girls making up 50.6% of the weighted sample, but the associated mathematics achievement scores highlight a gender imbalance with boys scoring significantly higher than girls in the G4 cohort. The analysis of variance revealed significant difference with mean mathematics achievement score for boys being higher than that reported for girls, $F(1,53815) = 175.9, p < 0.01$. The boys' mathematics scores reflect a wide range of achievement with both the lowest and highest achievement scores of the sample located in the boys' data set. In contrast, the mean science achievement scores for G4 boys and girls are not significantly different at the 1% level, $F(1,53830) = 5.459, p = 0.019$. The spread of achievement scores is again smaller for girls than boys, with girls having more clustered responses as shown in Table 4.1-1. This finding concurs with that reported by Powney (1996) who notes that males were more variable than females in test performance when dealing with aspects such as quantitative reasoning and spatial visualisation.

Table 4.1-1: AVERAGE achievement by GENDER

		Mathematics		Science	
GENDER	N	Mean	Std. Deviation	Mean	Std. Deviation
BOY	26,600	499.0	79.1	501.0	75.5
GIRL	27,200	490.4	70.9	499.6	67.2
Total	53,800	494.7	75.2	500.3	71.4

Figures have been rounded so columns may not equal total

In order to summarise distributions of achievement, the IEA identify five benchmark

cut points of 400, 475, 550, and 625 that are applied to the data to categorise achievement scores into bands of achievement from 'below 400' to 'at or above 625'. An analysis of gender broken down by 'band of achievement' shows the major discrepancies between girls' and boys' mathematics achievement scores occur in the upper tail of the achievement distribution as highlighted in Figure 4.1-2. Deary *et al.* (2003) report males in higher proportions at either extreme of the IQ distributions in their re-analysis of the Moray House Test, the main test used in the Scottish Mental Survey of 1932 (SMS_{1932}). In SMS_{1932} there was no gender difference in mean IQ scores but there was a more marked excess of boys at the extremes, with approximately 1.4 boys to every girl in the IQ 50 to < 60 and 130 to < 140 bands (Deary *et al.*, 2003: 538).

The current data shows a difference in achievement between boys and girls at both ends of the scale (Table 4.1-2: Gender distribution within achievement bands). At the lower and upper levels of achievement it is clear that boys feature more prominently, extending to 1.9 boys to every girl in the highest achievement band (n=1,908). There are also proportionally more boys (54%) in the second highest achievement band (n=11,032) as illustrated in Figure 4.1-1.

A similar profile is evidenced in the G4 Science data. Difference in achievement between boys and girls is concentrated at the extremes of the distribution, with 1.2 boys to every girl in the lowest achievement band (n=4,826) and over 1.8 boys to every girl in the highest achievement band (n=1,588), as illustrated in Figure 4.1-2.

Figure 4.1-1: Gender proportions within Mathematics Achievement bands

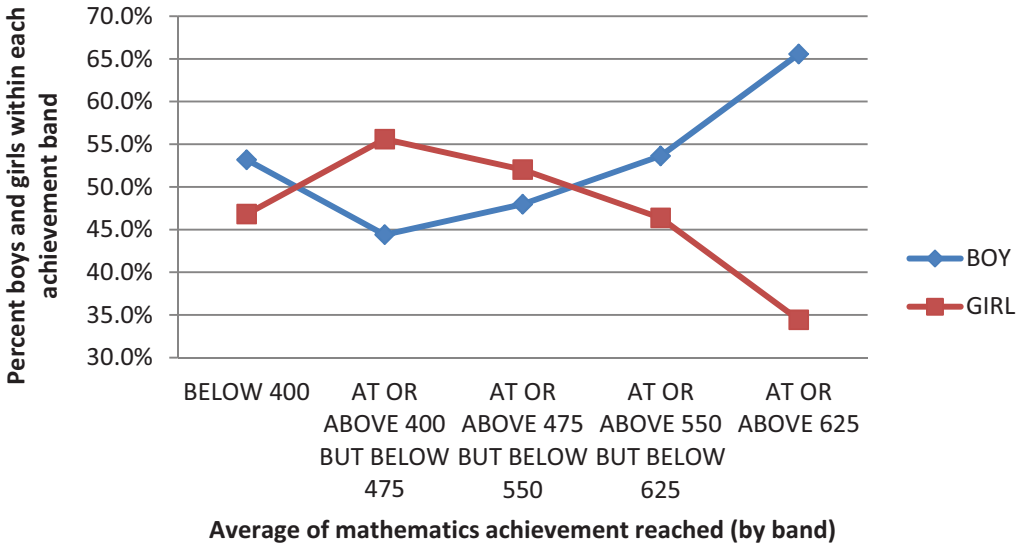
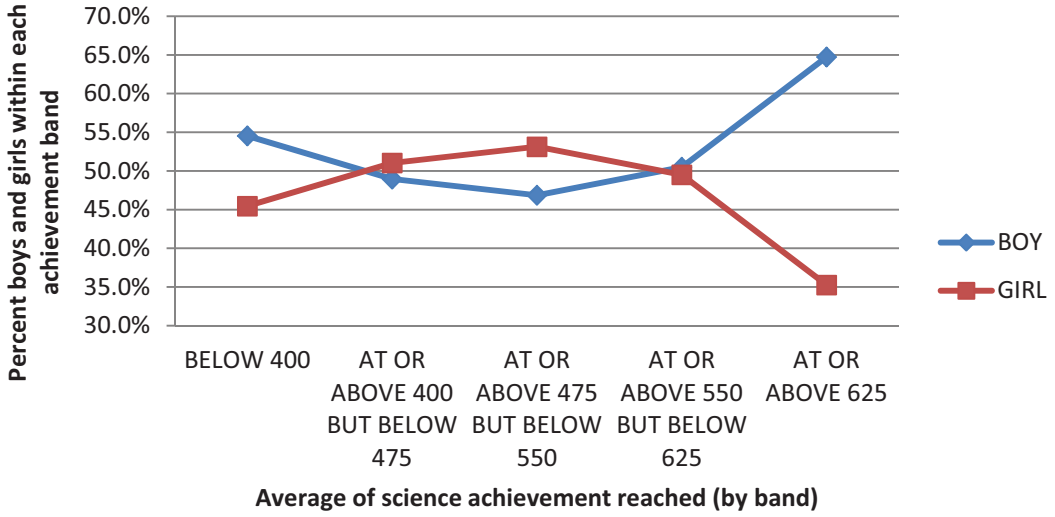


Figure 4.1-2: Gender proportions within Science Achievement bands



The major gender differences in the science data are confined to those extreme bands, with only smaller differences reported in the mid-range bands where achievement scores ranged from 400-625. These findings prompt the inclusion of 'gender' in subsequent analyses when considering associations between students' experiences and achievement scores.

Table 4.1-2: Gender distribution within achievement bands

	Mathematics			Science		
AVERAGE IBM_BANDS	BOY	GIRL	Total	BOY	GIRL	Total
BELOW 400	3,200	2,800	6,100	2,600	2,200	4,800
AT OR ABOVE 400 BUT BELOW 475	6,500	8,100	14,600	6,500	6,800	13,300
AT OR ABOVE 475 BUT BELOW 550	9,700	10,500	20,200	10,500	11,900	22,300
AT OR ABOVE 550 BUT BELOW 625	5,900	5,100	11,000	5,900	5,800	11,800
AT OR ABOVE 625	1,300	700	1,900	1,000	600	1,600
Total	26,600	27,200	53,800	26,600	27,200	53,800

Figures have been rounded so columns may not equal total

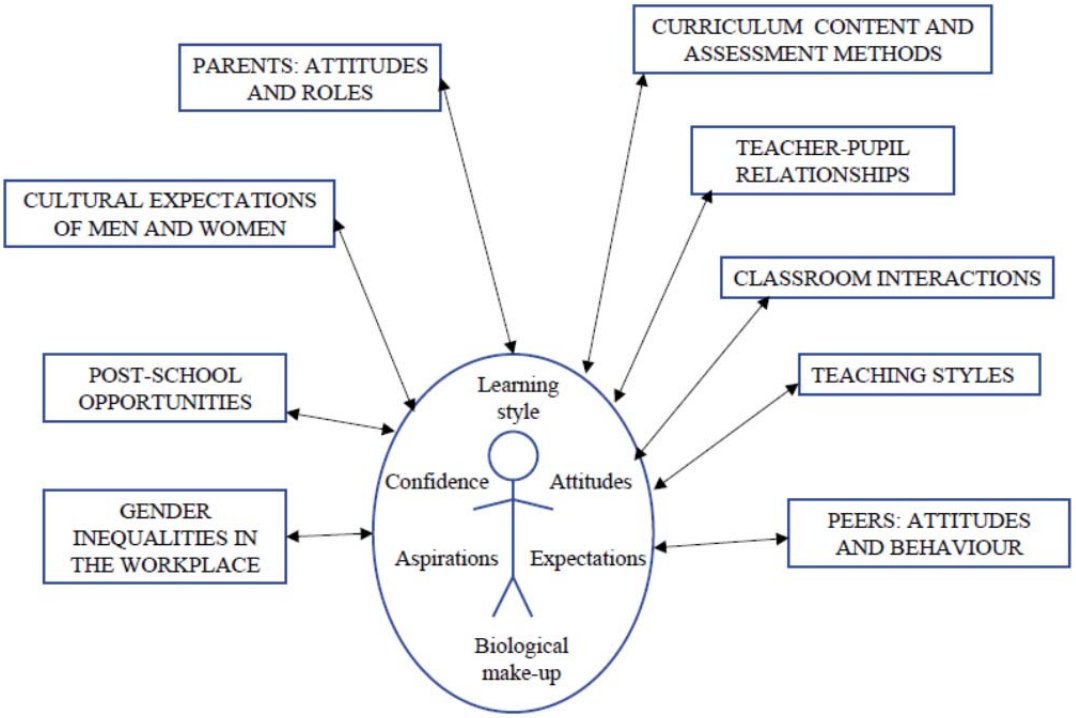
Recent surveys in Scotland (SSA, 2008; SSLN, 2012) suggest boys tend to have better mathematics attainment than girls across most stages and levels in Primary and lower Secondary years. Notably boys' attainment was better at the highest levels tested at each stage, while at the lowest levels tested at each stage girls tended to have higher attainment; this concurs in part with present data for G4 students. Although the age and stage of samples differ in $SSLN_{2011,}$ there is evidence of P4 and P7 boys outperforming girls, with about a five percentage point difference in the proportion of pupils performing 'well' or 'very well'

(SSLN, 2012). There was insufficient evidence of a difference in S2. For both boys and girls, there was an underlying pattern of decreasing attainment relevant to the expected level, with the largest drop noted between P7 and S2.

When considering science attainment of boys and girls separately, the underlying pattern of decreasing attainment, relative to expected levels, could be seen for both groups. In $SSA_{2007,}$ attainment levels in science *knowledge and understanding* showed a consistent gender difference in favour of the girls at Level A in P3, switching to a consistent difference in favour of boys at Level C and above in P5, P7 and S2; P5 and S2 being equivalent to G4 and G8 in current data. However, a similar analysis on science *literacy* items presents no consistent differences in achievement between girls and boys across stages and levels and in most cases the size of difference is small. At P7 the SSA reports relatively large differences with proportionally more girls than boys showing well-established or better skills at Levels C and D; where level D is the expected level of study for that stage.

Although Tinklin *et al.* (2001) note that in the past, when it comes to analysis of older students' performance in SQA examinations, males tended to out-perform females in certain subject areas such as mathematics and physics. However data from 1999 onwards show that female candidates at Standard Grade were more likely than males to gain awards at Credit level or at General/Credit level in almost every subject they entered, including mathematics but with the exception of general science. Similarly for performance in Higher grade there is a gender difference in favour of girls. Gender differences are well-documented at the secondary school stage, with the largest differences in performance found at the highest levels of attainment: more girls than boys gain five or more awards at 1-2 (Credit level) and 1-4 (General and Credit level), with a similar pattern found at Higher Grade (Tinklin *et al.*, 2001). In their review of literature on gender differences, Tinklin *et al.* (2001) outline a range of social, environmental and cultural factors that potentially influence gender differences between students' achievement, attitude, expectation, and confidence levels. Figure 4.1-3 displays the range of factors that emerged in that review, including the suggestion that a variety of different teaching strategies can be influential in maximising potential to engage with as many young people as possible. An emphasis in their research was to move away from any gender stereotyping, or viewing of boys and girls as homogeneous groups, to focus attention on improving classroom practice for all pupils.

Figure 4.1-3: Factors influencing gender differences (Tinklin *et al*., 2001: 7)



Geist and King (2008) contend it is important for teachers to consider actions that will support mathematics development in both boys and girls. They acknowledge there are differences in the way boys and girls learn, but argue due consideration is given to the impact of particular teaching practices on each gender and whether selected practices are of equal benefit to boys and girls. Some of the factors identified by Tinklin *et al.* (2001), in particular teaching styles, classroom interactions, and teacher-pupil relationships, are evidenced in the $TIMSS_{2007}$ survey data and provide an opportunity to evaluate the empirical data to ascertain whether associations between achievement and gender that can be attributed in part to the suggested factors. The title of Geist and King's article: 'Different, Not Better: Gender Differences in Mathematics Learning and Achievement' highlights the fact that various factors interact with gender but are not consistently directional; some factors associated with higher achievement for girls and others associated with lower achievement – differences in evidence, but not consistently favouring one gender.

4.1.2. Home background

A range of 'home' factors are now presented, taking account of each variable's distribution and association with achievement. A first cluster of variables takes account of parental background and language spoken at home, and a second cluster serves as a proxy for social and economic status of the household. For the first aspect, a new variable is

computed to combine data on students' place of birth and where they spent their early formative years. This variable provides data on whether or not a student was born in UK, and if not, whether they came to UK before school age (≤5 years). This is aggregated to a single dichotomous variable that summarises where they spent their early formative years. The standard naming convention for self-derived variables is: 'A' – G4, 'S' – Student, 'D' – Derived, with final letters used as a descriptor. In this particular case the derived variable is named ASDBFORM, a derived G4 student-variable with the last five letters used as a descriptor for: Born or spent early FORMative years in the country (BFORM)

Table 4.1-3: Formative years in UK (ASDBFORM)

BORN OR LIVED IN UK FOR FORMATIVE YEARS (≤5YRS)	Frequency	Percent
EARLY FORMATIVE YEARS SPENT IN UK	50,600	95.2
ARRIVED IN UK OLDER THAN 5	2,500	4.8
Total	53,200 54,000	100.0

Around 5% of respondents arrived in UK older than five years. The data in Table 4.1-4 show the feature of 'not having spent early formative years in the native country' has a pronounced association with mathematics achievement score; science achievement is similarly associated with this student variable but the overall achievement in science is higher, with more closely grouped data than witnessed within the mathematics scores.

Table 4.1-4: AVERAGE of achievement by Early Formative Years (ASDBFORM)

		Mathematics		Science	
BORN OR LIVED IN UK FOR FORMATIVE YEARS (<5YRS)	N	Mean	Std. Deviation	Mean	Std. Deviation
FORMATIVE YEARS SPENT IN UK	50,600	499.1	72.9	504.8	68.9
ARRIVED IN UK OLDER THAN 5	2,500	422.1	74.3	428.0	70.8
Total	53,200	495.4	74.8	501.1	71.0

Figures have been rounded so columns may not equal total

Another home-based variable concerns spoken language used at home, a measure of native language use at home with scope to provide potential indicator of bi-lingual status by reporting frequency of language skills in English being reinforced at home. The National Research Coordinator (NRC) decided on the 'native language' (multiple languages as necessary in some nations) to ensure test items were presented in the language used in

school. Students who could not read or speak the language of the test, and so could not overcome the language barrier of testing, were designated as non-native language speakers and could be withdrawn from the assessment; typically this excluded students who had less than one year of instruction in the language of the test (IEA, 2003). Students reported on frequency of using native language at home, picking the answer they 'think is best' from the options provided.

Fewer than 10% of students report either 'never' or only 'sometimes' speaking English at home, but the mean achievement scores for students in those categories is considerably lower than that reported by others. Some 13% of respondents report 'almost always' speaking English at home. Given the options open to the students, 'almost always' being selected instead of 'always' implies the student is in command of more than one language, which could be interpreted as an indication of bi-lingual status. Table 4.1-5 shows the students who identified themselves with the 'almost always' category show an association with high mean mathematics and science achievement scores and a narrow spread of achievement compared with those students who 'always' speak English at home. There also appears to be a sharp drop in achievement scores associated with the other categories of 'sometimes' or 'never', where English is not the dominant language used at home.

Table 4.1-5: AVERAGE of achievement by Own Language at Home (English)

			Mathematics		Science	
OWN LANGUAGE AT HOME	N	Percent	Mean	s.d.	Mean	s.d.
ALWAYS	42,100	78.4	495.3	74.9	501.5	71.3
ALMOST ALWAYS	7,000	13.0	517.0	71.9	520.0	67.6
SOMETIMES	3,200	6.0	466.2	65.6	468.5	59.9
NEVER	1,400	2.6	438.2	62.8	446.1	64.2
Total	53,800	100.0	494.7	75.2	500.3	71.4

A second cluster of 'home' variables draws on the availability of resources in the home environment. Home environments have been shown to be a major factor that influences overall child development, including early interactions with family members and quality of resources for learning that can impact on development (Iltus, 2006). Gottfried (1985) reviews broad measures of SES based on Duncan's *socioeconomic index*, Siegal's *prestige scale* and Hollingshead's *four factor index*, concluding that Hollingshead's four factor index is a strong candidate for informing child development research. This four factor index takes account of parental occupation, education, marital status and gender;

unfortunately these can be difficult measures to collect in large-scale surveys. It is not easy to produce sound measurements of the social and economic background of children, indeed it requires a substantial investment in data collection, coding, and data management (Hauser, 1994). For that reason, proxy measures for home influence on child development are used in the absence of better measures for parental background, such as parental income, parental education and parental occupation that are recognised as three main indicators of SES (Sirin, 2005; Mueller and Parcel, 1981).

The first environmental measure considered here is the 'number of books in the home', a proxy for parental interest and potential to engage in reading with the student as well as modelling study techniques and personal reading at home. Parental support can be through directly reading to their child or more implicitly by being active readers themselves, providing an interest in education through cultivating a learning environment at home where children are positively encouraged by observing their parents read: 'Parents who read frequently to their children are also likely to read more themselves, have more books (including children's books) in the home, take their young children to the library, and so on.' (Bus *et al.*, 1995). A number of studies have used this measure as a proxy for parental engagement and impact on students' educational attainment, including Byrne and Smyth (2010) who acknowledge: 'A large body of research in Ireland, as elsewhere, indicates significant differences in academic outcomes by aspects of parental background, including parental education, social class, number of books in the home and so on' (Byrne and Smyth, 2010: 13). The OECD includes this measure in their Programme for International Student Assessment (PISA) that reports on students aged around 15 years. Kirsch et al. (2002) attempt to quantify the impact of various measures on student development as captured in $PISA_{2000}$, reporting regression coefficients for a range of factors in relation to reading scores. These background factors, including 'number of books at home', explain variances within the data. Table 4.1-6 shows regression coefficients at the within-school level for cultural communication with parents, educational resources in the home and the number of books in the home.

Table 4.1-6: Multilevel regression coefficients for model of reading proficiency (extracted from Table 7.5 in Kirsch et al, 2002: 242)

	Cultural communication		Home Educational Resources		Number of Books	
	Regression coefficient	S.E.	Regression coefficient	S.E.	Regression coefficient	S.E.
Denmark	10.9	1.6	0.7	1.4	6.5	1.2
Finland	5.2	1.3	0.3	1.3	4.1	0.9
Ireland	1.7	1.4	4.8	1.5	6.7	1.1
New Zealand	-1.4	1.6	9.6	1.8	7.9	1.2
Norway	7.2	1.7	10.3	1.7	5.1	1.3
Sweden	6.9	1.4	1.3	1.2	5.5	1.1
United Kingdom	3.3	-1.3	3.6	1.3	5.6	1.0

The cultural communication between a student and his or her parents – discussing political or social issues; discussing books, films or television programmes; listening to classical music together – is associated with better reading performance in some countries, most markedly Denmark (10.9). In half of the countries, however, the regression coefficient for cultural communication is not significantly different from zero. The regression coefficients and standard errors in the final columns of Table 4.1-6, for the number of books in the home, show a fairly consistent and significant impact in the countries listed. The overall range in $PISA_{2000}$ runs from 0.6 in Brazil to 9.3 in the United States, and the regression coefficients are statistically different from zero in all except three countries: Brazil, Italy and Poland (Kirsch *et al.*,2002). Similarly, on the availability of educational resources in the home, such as text books, a desk, and a quiet place to study, a substantial impact is noted in some countries, including Norway (10.3) and New Zealand (9.6), but availability of educational resources appears to have no impact in a number of other countries such as Finland (0.3). A factor on 'home resources' will be discussed and developed for later inclusion in my models of TIMSS data.

Socio-economic status, insofar as value is placed on education and the number of possessions in the home related to education, is observed to provide an advantage to students where home support is high. Students were asked to estimate the number of books in their home using a guide of shelves or bookcases to help quantify the number of books, reported as shown in the frequency distribution below. Their responses were coded into five categories: 0 to 10 books; 11 to 25 books; 26 to 100 books; 101 to 200 books, and more than 200 books.

Just under one third of G4 students report having 25 or fewer books at home, a feature that is associated with low mean achievement scores as illustrated in Table 4.1-7 for science achievement data. Having 'none or very few' books appears to be associated with

lower mathematics achievement and lower science achievement.

Table 4.1-7: Average of mathematics and science achievement by number of books

NUMBER OF BOOKS AT HOME (AS4BOOKS)	N	Percent	Mathematics Mean	Mathematics s.d.	Science Mean	Science s.d.
0 TO 10 BOOKS	6,500	12.1	439.3	72.0	445.5	66.7
11 TO 25 BOOKS	10,500	19.7	474.3	68.4	480.2	63.5
26 TO 100 BOOKS	17,500	32.8	503.3	67.2	504.5	63.7
101 TO 200 BOOKS	9,900	18.6	519.0	67.6	528.1	62.7
OVER 200 BOOKS	9,000	16.8	517.9	79.0	527.6	74.8
Total	53,400	100.0	495.2	74.9	500.3	71.4

All three of the derived variables for 'own language', 'early formative years' and 'number of books', show a strong measure of association with mean mathematics and science achievement, each demonstrating a significant association at the 0.01 level (2-tailed) on Spearman's rho. The 'number of books at home' appears to be a particularly strong indicator variable given the strength of association with mathematics achievement ($\rho = 0.32$) and science achievement ($\rho = 0.34$) reported in this analysis; much as expected on the basis of findings reported by Kirsch (2002).

4.1.3. Home resources

A second measure of home support to be considered for analysis is a combination of responses on the availability of resources and facilities at home. This measure also serves as a proxy for parental support as discussed above. A Principal Component Analysis (PCA) was undertaken on the responses to the eight home resources identified in the student questionnaire – possession of calculator, computer, study desk, dictionary, internet connection, own bedroom, mobile phone and encyclopaedia. This analysis used Kaiser's Eigenvalue-one criterion and Catell's Scree graph to guide the selection of variables as recommended in Tinsley and Tinsley (1987). Responses to the 8-item section of the questionnaire were subjected to a PCA using ones as prior communality estimates. The principal axis method was used to extract the components, and this was followed by a varimax (orthogonal) rotation, to generate uncorrelated components. This rotation method retains orthogonal axes (perpendicular) but transforms the axes to maximise the variance explained by the derived components. The resultant components are clearer to interpret than in the un-rotated component matrix format, in that there will be high loadings on some variables and low loadings on others. The varimax rotation method maximises the dispersion of loadings across components, as it tries to load a smaller number of items onto

each component, resulting in more interpretable clusters of components (Field, 2000).

Using an oblique rotation method, such as Oblimin in SPSS, would result in a transformation of axes where they are no longer restricted to be perpendicular. If the components interact with one another then the axes will no longer be orthogonal, a condition that is problematic and limiting in terms of how those components can be used in subsequent analyses. Given the resultant components are ultimately going to be used in regression models to explain variance in achievement scores, they should as far as possible be uncorrelated; hence my use of varimax rotation to finalise factor loadings for the derived principal components.

In interpreting the rotated factor pattern, an item was taken to load on a given component if the factor loading was 0.4 or greater for that component, and was less than 0.4 for the other components. If an item appears to contribute to more than one component it is removed from the analysis since it will not be possible to assign uniquely to any one component, i.e. if it shows a loading of more than 0.4 on two or more components it cannot be claimed to make a unique contribution to one component. Using these criteria, the variable on possessing a 'study desk' was dropped from the analysis as it loaded on more than one component and also because it accounted for a low proportion of common variance (0.33) as documented in the extracted communalities following an initial exercise in dimension reduction.

Table 4.1-8: Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		Rotation Sums of Squared Loadings		
Comp	Total	% of Variance	Cumulative %	Total	Cumulative %	Total	% of Variance	Cumulative %
1	1.85	26.5	26.5	1.85	26.5	1.50	21.5	21.5
2	1.16	16.6	43.1	1.16	43.1	1.48	21.1	42.6
3	1.04	14.9	58.0	1.04	58.0	1.08	15.4	58.0
4	0.93	13.3	71.2	0.93	71.2	1.00	14.3	71.2
5	0.82	11.6	82.9					
6	0.69	9.8	92.7					
7	0.51	7.3	100.0					

Extraction Method: Principal Component Analysis.

In this particular analysis I extended the number of components by relaxing the eigenvalue-one criterion because of the low percentage of total variance explained, leading to identification of four components as documented in Table 4.1-8. Three items were found to load on the first component, which was subsequently labelled 'study tools' (calculator, dictionary and encyclopaedia). Two items loaded on the second component, which was

labelled the 'ICT' component (computer and internet connection). The resultant analysis reduced the data to four components that suitably summarised the responses under headings of: home possession of 'study tools' and 'ICT' facilities, and retaining 'mobile phone' and 'own bedroom' as stand-alone resources that reflect on SES and having a personal (study) space. The four components as identified as above are justified in Table 4.1-8: Total Variance Explained and Table 4.1-9: Rotated Component Matrixa (Home Resources).

Table 4.1-9: Rotated Component Matrix[a] (Home Resources)

	Component			
	1	2	3	4
POSSESS\CALCULATOR	0.70			
POSSESS\COMPUTER		0.87		
POSSESS\DICTIONARY	0.75			
POSSESS\INTERNET CONNECTION		0.84		
POSSESS\OWN BEDROOM				1.00
POSSESS\MOBILE PHONE			0.99	
POSSESS\ENCYCLOPAEDIA	0.65			

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
Factor loadings of less than 0.4 are not shown
a. Rotation converged in 4 iterations.

Components 3 and 4 are loaded on single items and so possession of 'own bedroom' and possession of 'mobile phone' are retained as separate variables. For components 1 and 2, I used factor-based scores rather than newly created variables based on the PCA factor scores (Cadima & Jolliffe, 1995). This resulted in derived variables for 'study tools' (component 1) and 'ICT' (component 2) based on mean responses to contributing items. This approach reduces the amount of data used, with approximate factor scores being calculated i.e. factor scores based on only a few variables that have the highest factor loadings (Gorsuch, 1983). Information is lost in this process because the approximate factor-based score strategy allows each item to contribute to only one component and equally weights all items assigned to a factor, whereas factor scores would take into account each variable's contribution to the component, however small. This approximation means the components will not be strictly uncorrelated, but it does make for a straightforward interpretation and subsequent re-calculation within other data sets.

Fewer than ten per cent of students had none or one of the study tools, whereas nearly seventy per cent had all three of calculator, dictionary and encyclopaedia at their disposal for home study. There is a fairly linear association between the number of study tools and mean achievement scores, with higher mean scores associated with having all three study tools at home as documented in Table 4.1-10.

Table 4.1-10: AVERAGE of achievement by Study Tools (calculator, dictionary, and encyclopaedia)

POSSESS\STUDY TOOLS (CALC, DICT, ENCYCL)	N	Percent	Mathematics Mean	s.d.	Science Mean	s.d.
NO STUDY TOOLS	1,200	2.3	420.9	79.6	427.6	80.4
ONE OF CALC, DICT, ENCYCLOPAEDIA	3,900	7.4	454.9	75.6	463.3	71.1
TWO OF CALC, DICT, ENCYCLOPAEDIA	12,100	23.0	482.3	74.1	487.6	69.5
ALL THREE OF CALC, DICT, ENCYCLOPAEDIA	35,500	67.4	507.7	70.3	512.9	67.2
Total	52,800	100.0	496.0	74.5	501.5	70.9

The majority of students (95%) report having a home computer. Within that group only 10% do not have an internet connection. A very small proportion of respondents have an internet connection but no home computer; presumably making use of the internet connection in association with a smart phone or equivalent handheld device, but we do not know from the data gathered. Computer ownership is clearly very high, noting responses excluded games packages such as 'PlayStation, GameCube, Xbox, or other TV/video game computers' ($TIMSS_{2007}$: G4 questionnaire). However, we do not know how these computers are used, whether in support of study skills or solely for recreational internet surfing or equivalent. We cannot say how the technology may or may not have a direct influence of their academic studies but the potential is there to be exploited; this measure also serves as an indicator of home support along the lines of 'number of books'. The empirical data shows possession of these home ICT resources is associated with mean mathematics and science achievement; the association of data on 'ICT capabilities at home' with achievement is illustrated in Table 4.1-11.

Table 4.1-11: AVERAGE of achievement by ICT (COMPUTER/INTERNET)

POSSESS\ICT (COMPUTER+/-INT)	N	Percent	Mathematics Mean	s.d.	Science Mean	s.d.
NO HOME COMPUTER OR INTERNET	2,500	4.8	449.8	80.3	449.8	76.2
INTERNET WITHOUT COMPUTER	500	0.9	431.3	90.0	445.5	87.2
COMPUTER WITHOUT INTERNET	5,100	9.5	457.0	76.8	462.5	73.6
HOME COMPUTER AND INTERNET	44,900	84.8	503.5	70.9	509.3	67.1
Total	53,000	100.0	495.9	74.4	501.5	70.9

There appears to be a marked difference in achievement scores when considering associations with ICT, in particular whether or not the student has an internet connection as

well as a home computer; possessing a home computer without internet functionality is associated with low achievement scores in both the mathematics and science cognitive items. Those divisions of possessing a computer with or without an internet connection will be explored in later analyses.

Possession of a mobile phone within the G4 student cohort is high, with over 85% reporting ownership. It is worth noting however that association with mathematics achievement is reversed, with significantly higher mean achievement scores associated with those who *do not* possess a mobile phone, as illustrated in Table 4.1-12. A high proportion of Primary 5 students report possession of a mobile phone, but unlike possession of other technologies, ownership does not appear to be associated with higher achievement in either mathematics or science.

Table 4.1-12: AVERAGE of mathematics achievement by Mobile Phone (AS4GTH07)

POSSESS			Mathematics		Science	
MOBILE PHONE	N	Percent	Mean	s.d.	Mean	s.d.
YES	46,000	85.8	492.1	74.6	497.3	70.5
NO	7,600	14.2	511.3	75.1	519.6	73.3
Total	53,600	100.0	494.8	75.0	500.4	71.3

The final variable in this cluster of home-based features addresses whether students have their 'own bedroom' or not. Just over three-quarters of the students report having their own bedroom (77%), a feature that has the potential to be associated with having a study space at home. Less than one fifth of those students who have their own bedroom indicate they do not have a 'study desk'; and two thirds of those without a 'study desk' have their own bedroom. Taking those proportions from Table 4.1-13 together, there are only around 6% of students who have neither a study desk nor their own bedroom.

Table 4.1-13: Cross-Tabulation of Own Bedroom & Study Desk

COUNT		POSSESS STUDY DESK		Total
		YES	NO	
POSSESS OWN	YES	33,800	7,300	41,000
BEDROOM	NO	9,000	3,300	12,300
Total		42,800	10,600	53,400

Responses to this question on possessing own bedroom, presents a proxy for having a study space at home; noting the earlier scratching of 'possessing a study desk' from the PCA because of the double loading with this component and 'study tools'. This variable on 'possessing own bedroom' could also serve as an indicator of SES. The empirical data

shows an association between having one's own bedroom and achievement scores in mathematics and science, where not having a personal space is associated with lower achievement scores.

Table 4.1-14: Average of achievement by Own Bedroom (AS4GTH06)

			Mathematics		Science	
POSSESS OWN BEDROOM	N	Percent	Mean	s.d.	Mean	s.d.
YES	41,200	76.7	499.0	73.6	505.2	70.0
NO	12,500	23.3	481.8	77.9	485.6	73.1
Total	53,600	100.0	495.0	74.9	500.6	71.2

All of these home-variables show a strong measure of association with mean mathematics achievement. The strongest associations with mathematics achievement are with 'study tools' and 'computer +internet' as shown in Table 4.1-15; all four of the derived variables demonstrate a significant association with achievement at the 0.01 level (2-tailed) on Spearman's rho.

Table 4.1-15: Correlations (Home)

Spearman's rho		AVERAGE OF MATHEMATICS ACHIEVEMENT PVS (N=53,900)
GEN\HOME POSSESS\ICT COMPUTER+INTERNET	Correlation Coefficient Sig. (2-tailed) N=52,900	0.22** 0.00
GEN\HOME POSSESS\STUDY TOOLS (CALC, DICT, ENCYCL)	Correlation Coefficient Sig. (2-tailed) N=52,700	0.23** 0.00
GEN\HOME POSSESS\MOBILE PHONE	Correlation Coefficient Sig. (2-tailed) N=53,500	0.10** 0.00
GEN\HOME POSSESS\OWN BEDROOM	Correlation Coefficient Sig. (2-tailed) N=53,500	-0.10** 0.00

** Correlation is significant at the 0.01 level (2-tailed).

4.1.4. Out-of-school interests and experiences

The eight out-of-school interests and experiences that students reported on were: watching TV or videos; playing computer games; playing or talking with friends; doing jobs at home; playing sports; reading books for enjoyment; using the internet; and doing homework. The response to how much time students spend on out-of-school activities was on a five point ordinal scale going from 'no time' through 'less than 1 hour', '1-2 hours', 'more than 2 hours but less than 4 hours' to '4 or more hours' in any normal school day.

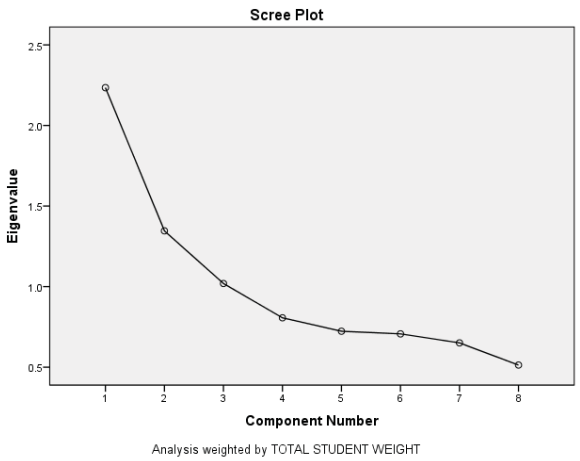
Responses to the 8 items were subjected to a principal component analysis using ones as prior communality estimates. The principal axis method was used to extract the components, and this was followed by a varimax (orthogonal) rotation as discussed for Home Resources.

Table 4.1-16: Total Variance Explained (Outside School)

	Initial Eigenvalues			Rotation Sums of Squared Loadings		
Comp.	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.24	27.9	27.9	1.72	21.8	21.8
2	1.35	16.8	44.8	1.51	18.8	40.6
3	1.02	12.7	57.5	1.35	16.9	57.5
4	0.81	10.1	67.6			
5	0.72	9.0	76.6			
6	0.71	8.8	85.4			
7	0.65	8.1	93.6			
8	0.52	6.4	100.0			

Extraction Method: Principal Component Analysis.

Figure 4.1-4: Scree Plot for Outside School



The first three components displayed eigenvalues greater than 1, generating the values presented in Table 4.1-16: Total Variance Explained (Outside School), and confirmed as appropriate through the Scree plot shown in Figure 4.1-4 where the elbow is used to signify selection cut off point. Combined, components 1, 2 and 3 accounted for 58% of the total variance.

Extra-curricular activities and out-of-school interests and experiences have been reduced to three components, structured around the headings: Individual (IND), Personal Development (PDEV), and Social (SOC). These groupings are supported by PCA as detailed in Table 4.1-17: Rotated Component Matrixa (Outside School), where the questionnaire items and corresponding factor loadings are presented.

Table 4.1-17: Rotated Component Matrix[a] (Outside School)

Things you do outside school, spend time …	Component		
	1 (IND)	2 (PDEV)	3 (SOC)
WATCH TV OR VIDEOS	0.77	0.07	-0.01
PLAY COMPUTER GAMES	0.82	-0.01	0.13
PLAY TALK WITH FRIENDS	0.16	0.09	0.76
DO JOBS AT HOME	-0.01	0.72	0.17
I PLAY SPORTS	0.08	0.09	0.82
READ BOOK FOR ENJOYMENT	0.03	0.74	-0.07
USE INTERNET	0.65	0.12	0.22
DO HOMEWORK	0.14	0.65	0.12

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 4 iterations.

As discussed above when extracting components within analysis of Home Resources, basing the components on approximate factor scores doesn't take full account of the data. Two alternatives are now considered in arriving at suitable components for inclusion in subsequent analyses of achievement data. First the components will be derived using Anderson-Rubin factor scores and second a rescaling mechanism is illustrated, picking up on the method as outlined in section 3.2.4 for the rescaling of ordinal polytomous variables.

As a primary method on this occasion, the components from Table 4.1-17 are generated using the Anderson-Rubin factor scores produced by PCA on SPSS, making the resultant components uncorrelated and therefore suitable for inclusion as independent explanatory factors in later regression models. These factor scores can be meaningfully interpreted given the minimal loadings contributed by items that are more firmly aligned with another component. For instance surfing the internet is primarily a solitary activity as it is hard to share a screen or co-browse. The end-user will usually be on their own, operating independently, even although part of this type of computer activity may involve using social networks. This first component is regarded as an 'individual' component (IND), highly loaded by 'watching TV', 'playing computer games' and 'using the internet', as shown in Table 4.1-17. Playing computer games and browsing the internet are therefore the dominant elements for the derived variable on individual activity (IND) accepting there may be a small social element (SOC) where those games are played with friends (loading of 0.16), or when a piece of collaborative homework (PDEV) is pursued on the internet (loading of 0.14).

The second component is about furthering personal development (PDEV), primarily though doing jobs at home, reading or doing homework, but this strand also has a small contribution from using the internet (loading of 0.12), a contribution that rests quite

comfortably within the overall experience and interpretation of furthering personal development. The dominant features of the third component are those of active and sociable activities (SOC), but this will also include spending time with friends on either a computer game, doing jobs at home, internet browsing or doing homework (loadings of 0.13, 0.17, 0.22 and 0.12 respectively).

These three derived variables are strongly statistically significant, but offer a substantively weak correlation with achievement scores as illustrated in Table 4.1-18.

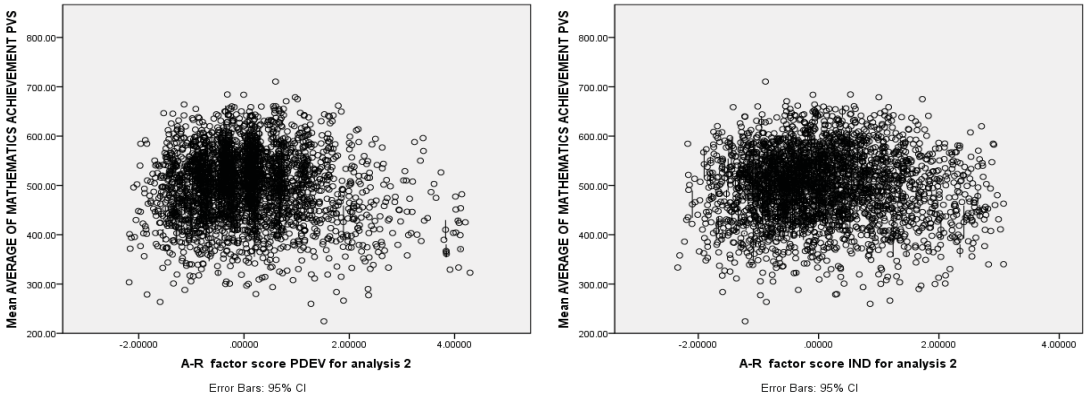
Table 4.1-18: Correlations of Anderson-Rubin factors with mean Maths Achievement Score

		A-R factor score IND	A-R factor score PDEV	A-R factor score SOC
AVERAGE OF MATHEMATICS ACHIEVEMENT PVS	Pearson Correlation	-0.043**	-0.077**	-0.013**
	Sig. (2-tailed)	0.000	0.000	0.006
	N	47,900	47,900	47,900

**. Correlation is significant at the 0.01 level (2-tailed).

They appear to have a weak non-linear association with achievement, with low and high levels of participation associated with lower achievement scores, as illustrated most noticeably with the 'Individual' and 'Personal Development' components presented in Table 4.1-19 and Figure 4.1-5.

Figure 4.1-5: Association between Anderson-Rubin factors and Mathematics Achievement



Using the Anderson-Rubin factor scores was an alternative to a factor-based scores method, as pursued with the Home Resources cluster outlined earlier. That approach will reduce the dimensionality of the data whilst retaining a meaningful construct that can be interpreted easily. For example, here the three derived variables would be generated on the basis of the average response recorded across the contributing items. The IEA use such a

technique within the published data, generating derived variables using the 'combination of responses method' to construct indices. Index scores are determined by classifying the cases into a high, medium, or low level of an index depending upon the combination of responses provided to the original items. For example:

> In the index of Good School and Class Attendance, cases were classified into the high index level if the three source items (arriving late at school, absenteeism, and skipping classes) were reported to be not a problem. Cases went to the low index level when two or more behaviors were reported to be a serious problem or two behaviors were reported to be a minor problem and the third a serious problem. The medium level included all other combinations of responses. In addition to constructing indices, the combination of responses method also was used to construct some specific derived variables. (TIMSS Technical Report, 2003:315)

Three ways of handling ordinal polytomous variables such as these are outlined and discussed in the methods chapter (3.2.6). The first option is to consider an IEA-style 'index' to partition the variance in the model; second option is to retain an extended categorical response for the variable, taking each value from the average of combined scores; and third option involves a rescaling of the derived metric to take account of the observed data, following Snell's scaling procedure for ordered categorical data (Snell, 1964).

By way of illustration, using the factor-based scores for Individual, Personal Development, and Social components from Table 4.1-17, the IEA-style indices can be computed as follows. Given the five point ordinal scale going from 'no time' through 'less than 1 hour', '1-2 hours', 'more than 2 hours but less than 4 hours' to '4 or more hours' in any normal school day, indices based on the IEA-style of classification are generated with LOW as an average of less than or equal to 2 (less than 1 hour); HIGH for 4 or more (more than 2 hours); and MEDIUM (MED) being those spending an average of between 1 and 2 hours on activities within the component. The distribution of Individual activities and associated achievement scores are presented in Table 4.1-19.

Similar patterns of association with achievement scores are in evidence across both disciplines, with highest achievement associated with MID-index of time on 'individual' activities and lower achievements associated with LOW and HIGH indices. The distinction between LOW and MID index is less pronounced when considering associations within science achievement data, whereas the HIGH index is clearly associated with low achievement scores in both disciplines.

Table 4.1-19: Index for Individual activity (ASDIND) - distribution and associated achievement scores (s.d.)

		Frequency	Percent	Valid Percent	Cumulative Percent	Mean Mathematics Achievement	Mean Science Achievement
Valid	LOW	17,700	32.8	33.0	33.0	490.5 (75.5)	501.6 (73.4)
	MID	30,300	56.1	56.5	89.4	502.5 (72.8)	504.8 (68.2)
	HIGH	5,700	10.5	10.6	100.0	467.5 (78.3)	474.0 (75.3)
	Total	53,600	99.4	100.0		494.8 (75.1)	500.5 (71.3)
Missing	System	300	0.6				
Total		54,000	100.0				

Comparable analyses of associations between achievement scores in mathematics and science and the other derived indices on out-of-school activities and experiences ('Social' and 'Personal Development'), are presented in Appendix 4. Similar patterns of association as noted above are in evidence, with highest achievement scores associated with mid-index engagement with the out-of-school-activities. Low levels of social activity are associated with low achievement scores, with a more pronounced association noted with mathematics achievement as opposed to science achievement.

In contrast to the Individual and Social variables, it is primarily only 'high' levels of reported engagement with Personal Development activities that are associated with low achievement scores. The core contributing elements to this derived variable are 'doing jobs at home', 'reading' or 'doing homework', activities that might be argued as supportive of learning. However, when these are in excess, with students spending an average of more than 2 hours per night on each of those aspects, there appears to be a negative association with achievement. This observation may reflect on the learners' *need* to do a lot of out-of-school study to keep up with, or to catch up on, work they have struggled with; it is worth noting however that only a small proportion of students align themselves in this category, with only 3.5 per cent of the G4 cohort reporting a HIGH index on the Personal Development derived variable.

Although these 'absolute' indicative measures of time spent on various out-of-school activities provide the analyst with relevant data, interpretation is not immediately transparent where the total amount of time that students self-report for those out-of-school activities will vary. For instance if each activity is reported as anything from no time to more than 4 hours, then on average across three activities a student attributed with a HIGH index (more than 2 hours on each activity), spends the equivalent of at least 6 hours per day on that cluster of activities.

An alternative representation is to consider a 'relative' measure, such as the proportion of time that students indicate for each of the clusters within their total out-of-school commitments. Students have their own 'total' based on the total allocation to activities outside school; technically scoring anything from 8 upwards (scores that equate to no time through to their total hours of out-of-school activity). What is of interest in the analysis is the way the students choose to use their time outside school, the relative time devoted to each cluster as a proportion of the total. Such an analysis will determine whether there are any associations between the proportion of time spent on different types of activity and the achievement scores recorded in mathematics and science. The proportion of time that students spend on each component is computed to give the variables with a PROP-prefix; with distributions as indicated in Table 4.1-20

Table 4.1-20: Proportion of time spent on each component of out-of-school activities (%)

	N	Minimum	Maximum	Mean	Std. Deviation
PROPIND	47,900	11	68	37.3	8.1
PROPSOC	47,900	8	56	30.7	7.3
PROPPDEV	47,900	11	69	32.0	7.7

By way of confirming the associations with achievement scores identified above, an index of these proportional distributions is derived. An equal division of time across these clusters would afford $33\frac{1}{3}$ per cent to each, so cut points based on 25 and 40 per cent are used to reflect 'low' and 'high' relative proportions – with an average of less than or equal to 25 per cent of their out-of-school activity taken as a 'low' proportion; 40 per cent or more taken as a 'high' proportion; and 'mid' being assigned to those spending an average of between 25 and 40 per cent of their total out-of-school time on the cluster of activities within that component.

Table 4.1-21: Average of mathematics/science achievement by index for proportion individual

		MATHEMATICS		SCIENCE	
INDEX FOR PROPIND	N	Mean	S.D.	Mean	S.D.
LOW (≤25%)	3,200	486.5	77.6	500.6	78.0
MID	26,900	500.2	73.4	506.1	70.5
HIGH (≥40%)	17,800	498.2	73.9	501.1	68.6
Total	47,900	498.5	73.9	503.9	70.4

Comparing the data presented in Table 4.1-21 with those presented in Table 4.1-19 , it is apparent that highest achievement scores continue to be associated with mid values in the index. However, the high index on the proportional variable (relative to other out-of-

school activities) shows a weaker association than that reported earlier on the absolute scale for a high take up of activities classified as 'individual'. Relatively low levels of engagement with 'individual' activities are associated with low achievement scores; a more noticeable association in mathematics than science.

A similar comparison with achievement data on both the 'social' and 'personal development' indices highlights an association of high achievement with mid values on the proportional index. Low achievement scores are associated with both low and high frequency of engagement with Social activities (Table 4.1-22) but a slightly different pattern emerges when considering responses concerning time spent on Personal Development (Table 4.1-21).

Table 4.1-22: Average of mathematics/science achievement by index for proportion social

		MATHEMATICS		SCIENCE	
INDEX FOR PROSOC	N	Mean	S.D.	Mean	S.D.
LOW (≤25%)	11,300	488.6	83.0	496.1	78.3
MID	31,000	502.6	71.1	507.8	68.7
HIGH (≥40%)	5,600	495.9	67.6	498.2	60.4
Total	47,900	498.5	73.9	503.9	70.4

Table 4.1-23: Average of mathematics/science achievement by index for proportion personal development

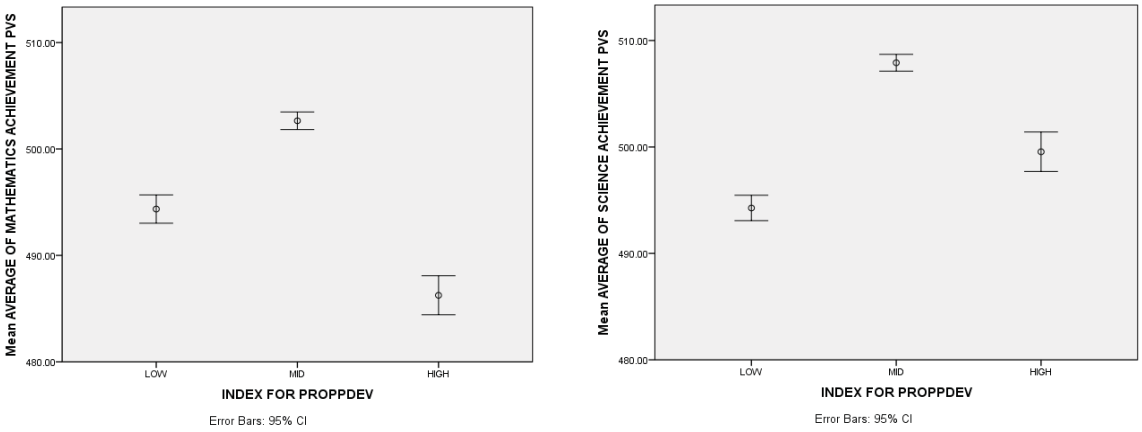
		MATHEMATICS		SCIENCE	
INDEX FOR PROPPDEV	N	Mean	S.D.	Mean	S.D.
LOW (≤25%)	9,700	494.4	66.8	494.3	60.2
MID	31,100	502.6	74.5	507.9	70.7
HIGH (≥40%)	7,100	486.3	78.7	499.5	79.7
Total	47,900	498.5	73.9	503.9	70.4

In contrast to the data in Table 4.1-22, a relatively low proportion of time spent on personal development is associated with low achievement scores (much as before) but the relatively high measure of engagement with 'doing jobs at home', 'reading' or 'doing homework' highlights a difference between disciplines in how those activities are associated with achievement (Figure 4.1-6). Students who report spending 40 per cent or more of their out-of-school activities on 'personal development' (high index) show an association with lower mathematics than science achievement scores.

Using the proportional measure takes account of the alternative options that students choose to engage with in out-of-school activities. The same broad picture is painted, with moderate levels of engagement associated with high levels of achievement and with Low and High levels of engagement more generally associated with low levels of achievement.

Using the IEA-style 'index' to partition the variance in the model simplifies the structure but in so doing ignores a lot of variance and potential explanation of association between these derived variables and the dependent variable on achievement.

Figure 4.1-6: Mean achievement scores by index of proportion personal development



The second approach set out in Chapter 3.2.6 for handling ordinal polytomous variables is to retain an extended categorical response for the variable, taking each value from the average of combined scores. For example, in the social component (ASDSOC) that combines 'play or talk with friends' (PLFD) and 'play sports' (PLSP), the original 5-point response with scores of 1 through 5 for 'no time' to '4 or more hours', extends to a 9-point scale when computing the average of the contributing scores (as shown in Table 4.1-24). For other components, where there are more contributing variables, the number of scale points will increase further. Taking this extended categorical scale as an ordinal measure for SOCIAL, there is no evidence to support a linear association of this derived variable with mean mathematics achievement; Spearman's Rho, of -0.004 does not support a linear association with mean mathematics achievement.

However, considering such an extended categorical scale as following an underlying continuous metric, a curvilinear relationship can be explored and analysed to determine whether this interval scale is associated with mathematics achievement.

Table 4.1-24: Derived scale for SOCIAL

ASDSOC	Frequency	Percent	Cumulative %
1.00	1,400	2.7	2.7
1.50	2,800	5.2	7.9
2.00	5,800	10.9	18.8
2.50	7,100	13.2	32.0
3.00	9,600	17.9	50.0
3.50	8,700	16.3	66.3
4.00	7,700	14.3	80.6
4.50	4,100	7.6	88.2
5.00	6,400	11.8	100.0
Total	53,600	100.0	

Figure 4.1-7 suggests the curvilinear association, which is supported by the Pearson correlation coefficients reported in Table 4.1-25 where the squared social variable (ASDSOCSQR) is highly correlated with mathematics achievement, showing significance at the 0.01 level (2-tailed).

Figure 4.1-7: Average of mathematics achievement – mean score (95% CI) and spread by SOCIAL

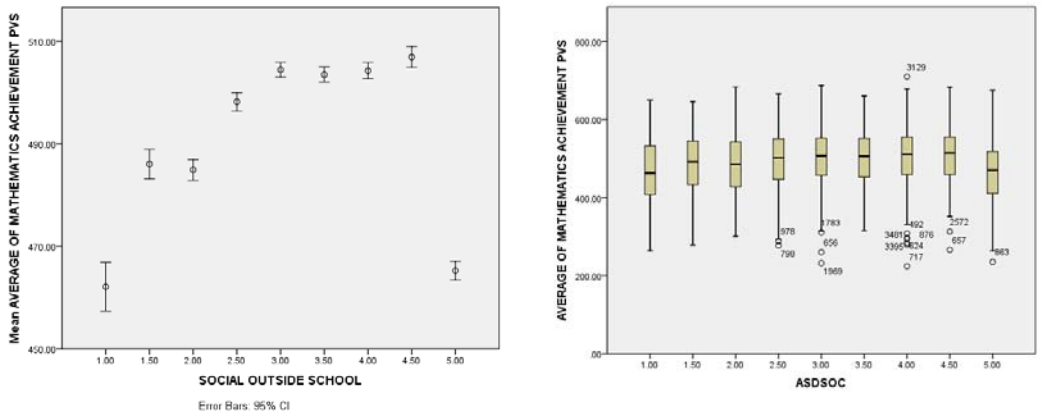


Table 4.1-25: Correlation with Mean Mathematics Achievement

		AVERAGE OF MATHEMATICS ACHIEVEMENT	SOCIAL OUTSIDE SCHOOL	SQUARED SOCIAL
AVERAGE OF MATHEMATICS ACHIEVEMENT PVS	Pearson Correlation	1	-0.003	-0.033**
	Sig. (2-tailed)		0.498	0.000
	N	53,960	53,610	53,610

** Correlation is significant at the 0.01 level (2-tailed).

A final development of scale involves a rescaling of the derived metric to take account of the observed data, following Snell's scaling procedure for ordered categorical data outlined in Chapter 3.2.6 (Snell, 1964). This is illustrated below, using the approximate

solution described in that chapter, and as applied to the SOCIAL variable considered above.

First, the empirical proportions in each category are used to replace the theoretical proportions $\hat{P}_{ij}$ in equations (3) and (4) replicated below.

$$0 = \frac{\partial L}{\partial x_j} = \frac{N_j}{\left[e^{(\hat{x}_j - \hat{x}_{j-1})} - 1\right]} - \frac{N_{j+1}}{\left[e^{(\hat{x}_{j+1} - \hat{x}_j)} - 1\right]} + N_j - \sum_{i=1}^{m} (n_{ij} + n_{i,j+1})\hat{P}_{ij}$$

$$\text{for } j = 2, \ldots k-2 \qquad (3)$$

$$0 = \frac{\partial L}{\partial x_j} = \frac{N_{k-1}}{[e^{(\hat{x}_{k-1} - \hat{x}_{k-2})} - 1]} + N_{k-1} - \sum_{i=1}^{m} (n_{i,k-1} + n_{ik})\hat{P}_{i,k-1}$$

$$\text{for } j = k-1 \qquad (4)$$

In this case, the scale has 5 categories ($k = 5$) and the distribution of responses for the contributing variables is as presented in Table 4.1-26. The corresponding data of observed proportions (p_{ij}) and cumulative proportions are provided in Table 4.1-27 and Table 4.1-28

Table 4.1-26: Distribution of responses SOCIAL

SOC	1	2	3	4	5	TOTAL
PLFD	4,309	12,788	12,112	9,569	13,897	52,675
PLSP	5,075	10,684	15,642	8,891	11,988	52,280
TOTAL(N)	9,384	23,472	27,754	18,460	25,885	104,955

Table 4.1-27: Observed proportions (p_{ij})

SOC	1	2	3	4	5
PLFD	0.082	0.243	0.230	0.182	0.264
PLSP	0.097	0.204	0.299	0.170	0.229
TOTAL(N)	0.089	0.224	0.265	0.176	0.247

Table 4.1-28: Observed Cumulative Proportions (p_{ij})

SOC	1	2	3	4	5
PLFD	0.082	0.325	0.555	0.736	1.000
PLSP	0.097	0.301	0.601	0.771	1.000
TOTAL(N)	0.089	0.313	0.578	0.753	1.000

The final set of data required to evaluate equations (3) and (4) are the values of n(i,j)+n(i,j+1) as presented in Table 4.1-29

Table 4.1-29: VALUES OF n(i,j)+n(i,j+1)

i\j	1	2	3	4	5
1		17,097	24,900	21,681	23,466
2		15,759	26,326	24,533	20,879

Starting with equation (4), where $k = 5$ & $j = 4$, the known values are substituted to give:

$$\frac{N_4}{[e^{(x_4-x_3)}-1]} = \sum_{i=1}^{2}(n_{1,4}+n_{1,5})p_{1,4} - N_4$$

$$\frac{N_4}{\sum_{i=1}^{2}(n_{1,4}+n_{1,5})p_{1,4} - N_4} = e^{(x_4-x_3)} - 1$$

$$\frac{18460}{21681 \times 0.736 + 24533 \times 0.771 - 18460} + 1 = e^{(x_4-x_3)}$$

Resulting in: $e^{(x_4-x_3)} = 2.125$

Following the same approach, substituting known values into equation (3) another two equations are formed:

$$e^{(x_3-x_2)} = 2.519$$

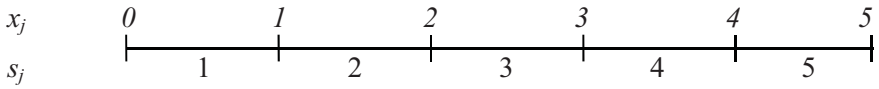
and

$$e^{(x_2-x_1)} = 5.601$$

Taking the natural logarithms of both sides and working from the basis of $x_1 = 0$, the values of x_j can be deduced to be: $x_2 = 1.723, x_3 = 2.647$ & $x_4 = 3.400$.

The final step is to evaluate the rescaled interval values. In a traditional scoring system the scale for five categories is developed from cut-points x_j (j= 0, 1, 2, 3, 4, 5) such that category s_j corresponds to the interval x_{j-1} to x_j

Figure 4.1-8: Intervals defining categories



As discussed in Chapter 3.2.6 the upper tail mean value is at a distance $-(log_e P)/Q$ beyond the point x_4, i.e. $-\frac{\ln(0.247)}{1-0.247} = 1.858$ beyond 3.400. Similarly for the average value in the lower tail, the mean value will be at a distance $-(log_e P)/Q$ below the point x_1 that has already been defined to be zero, i.e. $-\frac{\ln(0.089)}{1-0.089} = 2.651$ below zero. The mean values of the other boundaries are calculated to give the intermediate intervals for s. Those values and their Normalised equivalents (taking $\bar{x} = 1.735\ and\ \sigma = 2.619$) are presented in Table 4.1-30.

Table 4.1-30: Rescaled values for SOCIAL variables

j	1	2	3	4	5
s	-2.65	0.86	2.19	3.02	5.26
Normalised	-1.68	-0.33	0.17	0.49	1.35

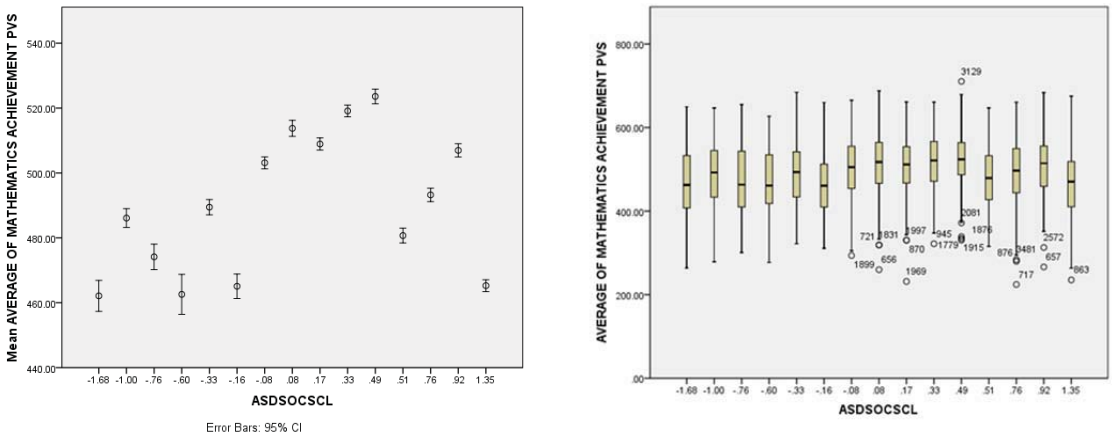
The G4 data is recoded by applying those values to the separate variables that contribute to SOCIAL, replacing the original scoring system of 1 to 5. The resultant responses are combined and averaged to give the distribution for a new derived variable for scaled social (SCLSOC) as presented in Table 4.1-31.

Table 4.1-31: Rescaled values for weighted SOCIAL (SCLSOC)

SCLSOC	Frequency	Percent	Cumulative Percent
-1.68	1,500	2.7	2.7
-1.00	2,800	5.2	7.9
-0.76	1,700	3.2	11.2
-0.60	900	1.6	12.8
-0.33	4,100	7.6	20.4
-0.16	1,300	2.5	22.9
-0.08	6,200	11.6	34.5
0.08	3,100	5.8	40.3
0.17	5,200	9.7	50.0
0.33	5,200	9.7	59.6
0.49	2,800	5.2	64.8
0.51	3,600	6.6	71.5
0.76	4,900	9.1	80.6
0.92	4,100	7.6	88.2
1.35	6,400	11.8	100.0
Total	53,600	100.0	

The association of the rescaled social variable (SCLSOC) with mean mathematics achievement is illustrated in Figure 4.1-9 . There is a significant association at the 0.05 level (2-tailed) between this scaled social variable and mathematics achievement scores; Spearman's rho correlation coefficient of -0.009. The squared variable is also strongly associated with mathematics achievement, significant at the 0.01 level (2-tailed); Spearman's rho correlation coefficient of -0.140, indicating a curvilinear association.

Figure 4.1-9: Average of achievement – mean score (95% CI) and distributional spread by scaled social (SCLSOC)



Although there is a lot of variability within responses a non-linear relationship can be seen with lower scores associated with the extreme values in the variable. As with the earlier extended categorical scale, this derived variable can be considered to follow an underlying continuous metric, which strengthens analysis as supported by Pearson's correlations shown in Table 4.1-32

Table 4.1-32: Correlations (Rescaled Social – linear and squared terms for scaled social variable)

		AVERAGE OF MATHEMATICS/ SCIENCE ACHIEVEMENT PVS	SCLSOC	SCLSOCSQD
AVERAGE OF MATHEMATICS ACHIEVEMENT PVS	Pearson Correlation	1.000	-0.009*	-0.140**
	Sig. (2-tailed)		0.037	0.000
	N	53,862	53,507	53,507
AVERAGE OF SCIENCE ACHIEVEMENT PVS	Pearson Correlation	1	-0.036**	-0.182**
	Sig. (2-tailed)		0.000	0.000
	N	53,975	53,625	53,625

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

Repeating the above for the other components derived from out-of-school activities gives rescaling values for 'individual' activities and those associated with 'personal development' as shown in Table 4.1-33.

Table 4.1-33: Rescaled values for INDIVIDUAL (IND) and PERSONAL DEVELOPMENT (PDEV)

	j	1	2	3	4	5
IND	Normalised	-1.56	-0.26	0.02	0.23	1.57
PDEV	Normalised	-1.44	-0.27	-0.05	0.08	1.68

A greater number of scale points are generated for those components, given the derived variables are based on three contributing elements. This illustrates the movement towards a continuous metric, an assumption that is made when rescaling the ordered categorical data using Snell's methods. Although there is wide variation within each category there appears to be a curvilinear association with achievement scores for both distributions, as illustrated in Figure 4.1-10: Distribution of achievement by scaled Individual and scaled Personal Development. The distribution of each derived variable is virtually identical for both disciplines, so only one is shown for illustration. This curvilinear association is supported as above by tabulating the Pearson correlation coefficients for association between mathematics (and science) achievement scores and the derived variables for the rescaled 'individual' activities (SCLIND & SCLINDSQD) as presented in Table 4.1-34.

Figure 4.1-10: Distribution of achievement by scaled Individual and scaled Personal Development

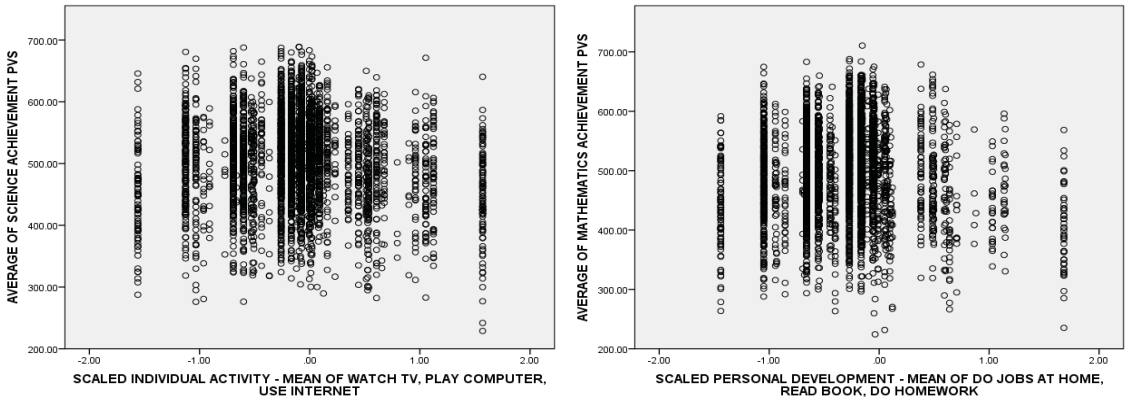


Table 4.1-34: Correlations (Rescaled Individual)

		AVERAGE OF MATHEMATICS/ SCIENCE ACHIEVEMENT	SCALED INDIVIDUAL (SCLIND)	SCALED INDIVIDUAL (SCLINDSQD)
AVERAGE OF MATHEMATICS ACHIEVEMENT PVS	Pearson Correlation	1	-0.051**	-0.213**
	Sig. (2-tailed)		0.000	0.000
	N	53,960	53,649	53,649
AVERAGE OF SCIENCE ACHIEVEMENT PVS	Pearson Correlation	1	-0.080**	-0.194**
	Sig. (2-tailed)		0.000	0.000
	N	53,975	53,664	53,664

**. Correlation is significant at the 0.01 level (2-tailed)

Similar results are shown for 'personal development' (SCLPDEV & SCLPDEVSQD) where support for a curvilinear association with achievement scores is confirmed though Pearson correlations presented in Table 4.1-35.

Table 4.1-35: Correlations (Personal Development)

		AVERAGE OF MATHEMATICS/ SCIENCE ACHIEVEMENT	SCALED PERSONAL DEVELOPMENT (SCLPDEV)	SCALED PERSONAL DEVELOPMENT (SCLPDEVSQD)
AVERAGE OF MATHEMATICS ACHIEVEMENT PVS	Pearson Correlation	1	-0.039**	-0.216**
	Sig. (2-tailed)		0.000	0.000
	N	53,960	53,637	53,637
AVERAGE OF SCIENCE ACHIEVEMENT PVS	Pearson Correlation	1	-0.015**	-0.245**
	Sig. (2-tailed)		0.000	0.000
	N	53,975	53,652	53,652

**. Correlation is significant at the 0.01 level (2-tailed)

The process of rescaling ordered categorical data using Snell's approximate method is shown to benefit analytical opportunities by increasing the power of resultant models. The observed association with achievement data is strengthened by basing the scoring system on the empirical findings, rescaling the metric in line with Snell (1964) and considering the derived variable to follow an underlying continuous measure. This makes the resultant derived variables more finely tuned to the data and will highlight potential associations with dependent variable. Where appropriate, this third development of rescaling ordered categorical measures will be used to investigate associations with achievement data in later multilevel regression models.

4.1.5. Classroom experiences (G4 mathematics)

Data considered in this section include learning experiences that reside within either the 'traditional' or 'reform' categories of practice discussed in the review of substantive issues. The reform movement calls for an increase in experiences that are claimed to enhance learning with a corresponding reduction in experiences that are of a traditional style and deemed to restrict students' understanding of content and achievement in mathematics and science. The experiences that come under the 'reform' category seek to provide a more effective education, boosting students' achievements and improving school effectiveness. Analyses of students' experiences and achievements will provide empirical evidence of the level of uptake within the disciplines in Grades 4 and 8, and insight to the relative effectiveness of those learning experiences and their association with achievements in mathematics and science. The empirical data will also address whether the 'reform' agenda is showing evidence of the claims made in the literature.

Respondents were asked to indicate the frequency of exposure to specified experiences or styles of learning, using a 4-point scale that runs from 'every or almost every lesson' through 'about half the lessons' and 'some lessons', down to 'never'. To ease final interpretation of these variables, responses on the scale were 'reversed' to present higher scores in line with the claimed benefit in learning and understanding, i.e. features that should be associated with high achievement scores. An exploratory data analysis of frequency of exposure was completed along with analyses of the mean mathematics and science achievement scores for each variable to confirm suitability for inclusion in subsequent multilevel models; analyses of experiences in science are presented separately from those in mathematics because the range of experiences is not replicated across disciplines or stages.

The selection of learning experiences identified in the literature review that are reported for students in G4 mathematics classes are:

Reform-based practice and Discussion

1. Explaining answers
2. Memorising formulae, processes and how to work out problems

Active Learning & Practical Activities

3. Measuring things in the classroom, taking a practical hands-on approach
4. Making tables charts or graphs, using data collected and having ownership of development and presentation of data

Learning Environment

5. I work by myself to answer questions
6. Working with other students in small groups

Contexts

7. Relating what is learned in mathematics to their daily life

The EDA confirms the suitability of including the majority of these variables. The distribution of responses and direction of association highlights where explanation of variance can be investigated through the multilevel models. In some instances, such as response on use of *contexts*, where there is minimal difference in achievement scores across response categories, the variable is unlikely to provide significant explanation of variance in the full model; this will be confirmed in subsequent analyses.

4.1.6. Classroom experiences (G4 science)

As with the classroom experiences in mathematics, the data under consideration includes learning experiences that reside within either the 'traditional' or 'reform' categories of practice discussed in science education. Respondents were asked to indicate the frequency of exposure to the experience or style of learning, using a 4-point scale that runs from 'at least once a week' through 'once or twice a month' and 'a few times a year', down to 'never'. The categories are much broader than those used for mathematics, presumably because science teaching in G4 is not necessarily explicitly pursued with the same regularity witnessed in mathematics education. As with the mathematics scales, to ease final interpretation of these variables, responses were 'reversed' to present higher scores in line with the claimed benefit in learning and understanding that should be associated with high achievement scores. An exploratory data analysis for each variable was undertaken to confirm suitability for inclusion in subsequent multilevel models; analyses of experiences in science are also presented in Appendix 4.

The selections of learning experiences identified in the literature review for students in G4 science classes are reported under the same broad headings:

Reform-based practice and Discussion

1. Write or give an explanation for something I am studying in science
2. Memorizing science facts

Active Learning & Practical Experiments

3. Watch the teacher do a science experiment
4. Design or plan a science experiment or investigation
5. Do a science experiment or investigation

Learning Environment

6. Work by myself to answer questions
7. Work with other students in small group on a science experiment or investigation

Contexts

8. Relate what is learned in science to their daily life

The EDA again confirms the suitability of including the majority of these variables. The distribution of responses and direction of association highlights where explanation of variance can be investigated through the multilevel models. As noted in the mathematics data, instances where there is minimal difference in achievement scores across response categories will provide limited explanation of variance; the effect of those variables will be confirmed in subsequent analyses.

In the reform movement there is a call for a reduction in 'memorising things' and passive learning, in favour of an increase in active learning and ownership of learning experiences. Data on students' engagement with practical aspects of science and opportunities for them to experience the scientific method will prove useful in evaluating levels of ownership and active learning that come through the cluster of variables 3 to 5 above, reporting on opportunities to watch, plan and do experiments or investigations. A distinction is drawn between opportunities to 'watch the teacher doing the activity' and students actively 'planning' the experiment or 'doing the experiment or investigation' themselves.

In all three situations a common finding was an association between low science achievement scores and students 'never' having opportunities to engage with science experiments or investigations. There are strong correlations between the three variables, as documented in Table 4.1-36, reflecting the possibility that the same students are affected in the majority of cases.

Table 4.1-36: Correlations (WATCH/PLAN/DO science experiments & investigations)

		WATCH TEACHER DO SCI EXP	PLAN SCI EXP OR INVEST	DO SCI EXP OR INVEST
WATCH TEACHER DO SCIENCE EXPERIMENT	Pearson Correlation		0.385**	0.447**
	Sig. (2-tailed)		0.000	0.000
	N=53,200		52,200	52,300
PLAN SCI EXPERIMENT OR INVESTIGATION	Pearson Correlation	0.385**		0.531**
	Sig. (2-tailed)	0.000		0.000
	N=52,600	52,200		51,800
DO SCI EXPERIMENT OR INVESTIGATION	Pearson Correlation	0.447**	0.531**	
	Sig. (2-tailed)	0.000	0.000	
	N=52,700	52,300	51,800	

**. Correlation is significant at the 0.01 level (2-tailed).

However, cross tabulations indicate otherwise. Focusing on the 'never' category, the response that is associated with low achievement scores, the partial breakdown of watching and planning science experiments for those who never do science experiments or investigations is presented in Table 4.1-37. A high proportion of over one fifth of students report never doing science experiments or investigations, a response that appears to be associated with low science achievement scores. If doing practical experiments or investigations is valued as suggested by the reform movement then this analysis provides data on whether there are opportunities for watching or planning when students never do experiments themselves.

Table 4.1-37: Analysis of 'never' DO experiments by PLAN and WATCH science experiment

NEVER DO SCI EXP OR INVEST	AT LEAST ONCE A WEEK	ONCE OR TWICE A MONTH	A FEW TIMES A YEAR	NEVER	Total
PLAN SCIENCE EXPERIMENT OR INVESTIGATION	500 (4%)	500 (4%)	2,000 (18%)	8,400 (74%)	11,400
WATCH TEACHER DO SCI EXPERIMENT	2,000 (18%)	1,700 (15%)	2,900 (26%)	4,800 (42%)	11,400

This breakdown shows that just fewer than three quarters of students who never do experiments report never planning investigations, but that nearly 60 per cent of those students watch a teacher doing experiments at least a few times a year. The combination of doing, watching and planning will be explored in the multilevel analyses, noting that opportunities for doing and planning are more closely aligned with a reform agenda, but that watching teachers do experiments may be the minimum expectation for engagement with practical activities and scientific method.

4.1.7. ICT experiences

ICT experiences in school, at home and in public spaces is reported to give a profile of availability and access. This wider information, given its direct focus on supporting schoolwork, complements that reported under 'home resources'. The distribution of response and associated achievement scores are presented in Table 4.1-38. A significant majority of students make use of computers at both home and school, a feature that is associated with high achievement in both disciplines. There is a small amount of variation in achievement scores dependent on whether students make use of computers in 'one' of home, school or other (such as a public library). Although less than one per cent of students do not use a computer at all, this lack of opportunity of experience is associated with a low mathematics achievement score and an even lower science achievement score in contrast to other response categories in science.

Table 4.1-38: Average of G4 mathematics/science achievement by computer use (home, school, other)

COMPUTER USE			Mathematics		Science	
	N	Percent	Mean	S.D.	Mean	S.D.
PC BOTH AT HOME AND AT SCHOOL	42,500	79.4	501.7	72.9	506.1	69.4
PC AT HOME BUT NOT AT SCHOOL	6,600	12.4	474.1	78.2	487.7	74.6
PC AT SCHOOL BUT NOT AT HOME	3,500	6.6	462.4	73.7	467.5	71.2
PC ONLY AT PLACES OTHER THAN HOME AND SCHOOL	400	0.8	471.8	62.4	480.4	48.4
DO NOT USE PC AT ALL	400	0.8	475.3	91.7	463.0	85.9
Total	53,500	100.0	495.2	74.8	500.8	71.1

A distinction between calculator and computer use is made within mathematics education. The reported frequencies and associated achievement scores are presented in Appendix 4. Around one third of students report 'never' using a calculator in their mathematics lessons, a feature that is associated with lower achievement scores than those associated with using a calculator in 'some lessons'. A high proportion of students report 'never' using a computer in mathematics classes – nearly 50 per cent of the cohort – with just over a further third using a computer in some of their lessons. More than half the cohort never uses a computer in science lessons with just over a fifth using a computer in some lessons. As with the mathematics data there is no obvious association between frequency of experience and science achievement scores, other than noting that the achievement scores dip with very frequent use of computers in lessons.

Student's use of computers in support of schoolwork, in and out of school, highlights highest achievement scores are associated with fairly infrequent use, and that high use of computers does not appear to offer any advantages in terms of achievement score. Those who 'never' use computers for their mathematics schoolwork, in or out of school, have a higher than average achievement score. The data indicates home computers are not used extensively to directly support school work whether internet connected or not.

4.1.8. School culture and ethos

Measures of school culture and ethos are derived from the two clusters of variables that relate first to students' perception of (i) peers' attitudes to study and (ii) teachers' expectation of students, and second how welcoming and 'safe' they find the school environment.

The climate of learning is reported through two variables that seek response on the extent to which: 'I think that children at my school try to do their best'; and 'I think that teachers at my school want children to do their best'. Responses for both items range from 'disagree a lot' through to 'agree a lot'. The climate of learning is taken to be the combination of these two variables, but given the skewed nature of the separate distributions a dichotomous derived variable is generated to reflect agreement with both variables ('a little' or 'a lot') or disagreement in one. This measure provides very limited evidence of variance across response categories.

The second measure of school ethos reflects how 'safe' students feel at school, taking into account their responses to five features of school life. An IEA derived variable produced an index of High, Medium and Low perception of being safe, based on the students' responses to the following set of questions:
In school, did any of these things happen during the last month?

1. Something of mine was stolen (AS4GSTOL);
2. I was hit or hurt by other student(s) (e.g., shoving, hitting, kicking) (AS4GHURT);
3. I was made to do things that I didn't want to do by other students (AS4GMADE);
4. I was made fun of or called names (AS4GMFUN);
5. I was left out of activities by other students (AS4GLEFT).

A high perception of safety was assigned on the basis of responding 'no' to all five statements; a low perception of safety assigned if there were three or more responses of 'yes'; with all other combinations falling into the medium category for perception of safety; the index was coded as missing if there were 2 or more source questions with invalid data. The data show an association between feeling safe in school and achievement, with lower achievement scores when students report a low perception of safety.

4.2. Student Questionnaire (G8)

A comparable exploration of the data from students in G8 was carried out. In the following sections any similarities and differences in response from G4 are highlighted. Where appropriate, further explanation or comment is provided but in the main, the bulk of findings documented in Appendix 4 lead to similar conclusions and justification for using these variables in subsequent analyses.

4.2.1. Biographical data

An even gender balance is reported in G8 but the associated mathematics and science achievement scores highlight a gender imbalance with boys scoring significantly higher than girls in the same stage. This finding concurs with research evidenced discussed in 4.1.1 (Powney, 1996 and SSLN, 2012).

The G8 variable that takes account of parental background is an extension of that presented within G4 analyses with an additional category to cater for those students who came to UK when older than 10 years. The three categories provide data on whether or not a student was born in UK, and if not, whether they came to UK before school age (≤5 years), during their primary education phase (aged 5 to 10 years), or when older than 10 years. The standard naming convention for G8 variables (as deployed by IEA) is to have the prefix 'B' – in this particular case the derived variable is BSDBFORM, a G8 student-variable derived from other registered variables, with the same naming convention as described earlier.

The data in Table 4.2-1 show the feature of 'not having spent early formative years in the native country' has a strong association with mathematics and science achievement score. As with earlier tabulated data, figures have been rounded so columns may not equal totals. A clear pattern of lower achievement scores appears to be associated with students who arrived in the UK in their later years, with those arriving after age 10 being associated with the lowest achievement scores in mathematics and science.

Table 4.2-1: AVERAGE of achievement by Early Formative Years (BSDBFORM)

		Mathematics		Science	
BORN OR LIVED IN UK FOR FORMATIVE YEARS (<5YRS)	N	Mean	Std. Deviation	Mean	Std. Deviation
FORMATIVE YEARS SPENT IN UK	55,900	490.5	75.5	498.7	75.9
ARRIVED IN UK AGED 5 TO 10 YEARS	700	452.3	90.3	459.6	93.9
ARRIVED IN UK OLDER THAN 10 YEARS	1,200	436.1	94.3	439.2	92.5
Total	57,800	488.9	76.6	497.1	77.1

The home-based variable concerning spoken language used at home (BS4OLANG) is reported with associated mean achievement scores in G8 mathematics and science. Unlike the data from the G4 students, there is no enhancement within the 'almost always' category, but rather a pattern of declining achievement associated with reducing levels of English spoken at home.

The other home-based variables, or environmental measures to be considered for inclusion are the same as in G4 with the 'number of books in the home', and 'home resources' available to students in support of their studies. The latter was broken down into four components following a PCA analysis of the data as detailed in Appendix 4, resulting in 'study tools' (calculator, dictionary, and encyclopaedia), ICT resources (computer and internet), possession of 'mobile phone', and 'own bedroom'. As with the G4 data, possessing a 'study desk' was dropped from the analysis because it was loaded on more than one component, following the same line of argument set out in 4.1.3. There is a linear association between the number of study tools and mean achievement scores in G8, with higher mean scores associated with having all three study tools at home.

A significant majority of the G8 students (97.5%) report having a home computer. Within that group only 6% do not have an internet connection. This measure is taken as an indicator of home support along the lines of 'number of books'. The empirical data shows possession of a computer *with* internet connection is positively associated with mean mathematics and science achievement. As with the other home resources, the effect of increased support in the form of ICT has a stronger association with science achievement scores.

A substantial majority of the G8 students report having their own bedroom (85%), a feature that has the potential to be associated with having a study space at home. The empirical data shows an association between having one's own bedroom and both mathematics and science achievement scores; a stronger association appears to be present within the science achievement data.

Possession of a mobile phone within the G4 student cohort was high, with over 85% reporting ownership. This rises to 97% of G8 students, a feature that appears to be associated with a small but significant *positive* effect on student achievement ($F=9.7$; $p=0.002$).

4.2.2. Out-of-school interests and experiences (G8)

A similar analysis of the out-of-school interests and experiences for students in G8 produced a different pattern of relationship. Responses to the nine items were subjected to a principal component analysis much as outlined in 4.1.4, using ones as prior communality estimates. The principal axis method was used to extract the components, and this was followed by a varimax (orthogonal) rotation as discussed earlier.

The resultant components are slightly different to those reported for G4 students, but still reflect on out-of-school activities as primarily: of an 'individual' nature (IND); as a reflection on 'personal development' (PDEV); and as a measure of 'social' engagement with other students (SOC); plus two standalone components that loaded on a single variable, i.e. Play Sport (PLSP) and Paid Work (WKPJ)

The 'social' dimension with this age group identified two variables as contributors for the component: play and talk with friends; and using the internet. This represents a shift in loading from that reported for G4 students, where 'using the internet' was aligned with solitary activities and grouped within the 'individual' component. For the older students the data shows a higher loading on the 'social' component, with likely links to social media on the internet (such as Facebook); I note in Table 4.2-2 that there remains a small loading on the 'individual' component (0.28) but that this loading is not included in subsequent analyses (dropped from consideration on ground of loading being less than 0.4).

Table 4.2-2: Rotated Component Matrix[a] (G8 Outside School)

Things you do outside school	Component		
	1 (IND)	2 (PDEV)	3 (SOC)
GEN\SPEND TIME\WATCH TV OR VIDEOS	0.80		
GEN\SPEND TIME\PLAY COMPUTER GAMES	0.85		
GEN\SPEND TIME\PLAY TALK WITH FRIENDS			0.86
GEN\SPEND TIME\DO JOBS AT HOME		0.66	
GEN\SPEND TIME\READ BOOK FOR ENJOYMENT		0.73	
GEN\SPEND TIME\USE INTERNET	(0.28)		0.71
GEN\SPEND TIME\DO HOMEWORK		0.76	

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
Factor loadings of less than 0.4 are not shown
b. Rotation converged in 4 iterations.

These three components for 'individual', 'personal development' and 'social' aspects of out-of-school activities are re-scaled in line with arguments presented earlier

(Snell, 1964). Table 4.2-3 presents the re-scaling factors for the score values contributing to each component.

Table 4.2-3: Re-scale factors for INDIVIDUAL (IND), PERSONAL DEVELOPMENT (PDEV) and SOCIAL (SOC) components

	j	1	2	3	4	5
IND	Normalised	-1.53	-0.47	0.04	0.46	1.50
PDEV	Normalised	-1.41	-0.56	-0.01	0.37	1.61
SOC	Normalised	-1.69	-0.34	0.17	0.54	1.31

The resultant associations with mathematics and science achievement scores support a curvilinear association with achievement scores for both distributions. When these variables are entered into the multilevel models the best fitting polynomial will be selected to maximise potential association with the dependent variable.

The two stand-alone components for 'playing sport' and having a 'paid job' were reported on the basis of time spent each school day, ranging from no time through to 4 or more hours. Just under a third of respondents indicate less than an hour or no time playing sport, with almost a fifth engaged in sporting activities for 4 or more hours each day. There is a distinct difference in uptake across genders (percentage of male /female participation rates are included in Table 4.2-4) but there are comparable numbers of boys and girls in the response categories that appear to offer the greatest association with enhanced achievement scores i.e. 1 to 2 hours and the neighbouring category of more than 2 but less than 4 hours. The empirical data supports the argument for an active schools agenda, encouraging participation in a broad definition of sport such as that offered by Council of Europe (2001):

> Sport means all forms of physical activity which, through casual and organised participation, aim at expressing or improving physical fitness and mental well-being, forming social relationships or obtaining results in competition at all levels.

Coulter (2005) notes there is no definitive evidence of a positive, causal relationship between physical activity/sport and academic achievement and although there are some studies where correlations have been found (Thomas *et al.* , 1994; Etnier *et al.*, 1997), the explanation for the nature and direction of association remains speculative. Coulter concludes that although the evidence is inconclusive on improved academic performance, there are indications that it does not have a negative effect and sporting activity does offer physical and emotional benefits. The empirical data here is less supportive of that final claim if high levels of participation are witnessed, where a negative association is documented when playing sport for 4 or more hours per day. However, participation in sporting activities for anything from 1 to 4 hours appears to be associated with higher levels

of achievement as set out in Table 4.2-4.

Table 4.2-4: Average of achievement by 'playing sport'

SPEND TIME\I PLAY SPORTS			Gender		Mathematics		Science	
	N	Percent	M(%)	F(%)	Mean	s.d.	Mean	s.d.
NO TIME	7,900	13.7	7.7	19.5	472.0	80.5	483.6	83.0
LESS THAN I HOUR	10,900	19.0	13.0	24.8	487.5	75.5	494.7	77.3
1 TO 2 HOURS	16,600	28.8	25.9	31.6	503.2	75.5	510.3	75.9
MORE THAN 2 BUT LESS THAN 4 HOURS	10,700	18.6	23.1	14.3	501.4	73.1	511.4	73.1
4 OR MORE HOURS	11,400	19.8	30.3	9.8	470.1	72.2	477.7	70.5
Total	57,400	100.0	100.0	100.0	489.0	76.4	497.4	76.9

Over three-quarters of the students do not work in a paid capacity. Of those that do, there appears to be a negative association with achievement scores in both mathematics and science as the level of paid work increases.

4.2.3. Classroom experiences (G8 mathematics)

Respondents were asked to indicate the frequency of exposure to the experience or style of learning, using a 4-point scale that runs from 'every or almost every lesson' through 'about half the lessons' and 'some lessons', down to 'never'. Responses on the scale were 'reversed' to present higher scores in line with the claimed benefit in learning and understanding. The selection of learning experiences identified in the literature review that G8 mathematics students report on, are:

Reform-based practice and discussion

1. Explaining answers
2. Memorising formulae and procedures

Active learning and practical activities

3. Deciding on own procedures for solving complex problems
4. Interpreting data in tables, charts or graphs

Learning environment

5. Work problems on our own
6. Working together in small groups
7. Listen to the teacher give a lecture-style presentation

Contexts

8. Relating what is learned in mathematics to their daily life

Technology

9. We use calculators
10. We use computers
11. Use a computer for mathematics schoolwork (in and out of school)

Assessment and feedback

12. Review our homework
13. Have a quiz or test

An exploratory analysis of exposure is presented in Appendix 4 for each experience in turn, confirming suitability for inclusion in later multilevel models. The mean mathematics achievement scores for each variable are presented separately from experiences in science because the range of experiences is not replicated across disciplines.

Some notable findings that can be followed through in the multilevel analyses include:

Solving complex problems

The association between levels of ownership of procedures for solving complex problems with achievement scores shows a marked reduction in achievement as the frequency of opportunity increases. This empirical finding is at odds with claims on the benefits of autonomy and ownership of strategies used within problem solving, where students who frequently decide on their own procedures for solving complex problems are associated with lower achievement scores.

Daily life

Just under one fifth of students never experience relating what they learn in mathematics to their daily life, but this does not seem to adversely affect their learning, in that the 'never' category is associated with a higher than average achievement score. The absolute difference in achievement scores is not as extreme as witnessed with some of the other variables, but the differences between groups are still statistically significant ($F_{3, 57912} = 418.1$, $p < 0.001$).

Lecture-style presentation

A high proportion of G8 students report listening to the teacher giving a lecture-style presentation on 'every or nearly every lesson'. This experience is associated with a high achievement score, whereas the lowest mathematics achievement score is associated with the 'never' category. Again there is a highly significant variance across categories where decreasing achievement scores are associated with the lower frequencies of a lecture-style of presentation ($F_{3, 58203} = 1113.5$, $p < 0.001$).

4.2.4. Classroom experiences (G8 science)

As with the classroom experiences in mathematics, the G8 data considered here relate to the predictor variables identified in Chapter 2. Respondents were asked to indicate the frequency of exposure to the experience or style of learning using the categories of 'every or almost every lesson' through 'about half the lessons' and 'some lessons', down to 'never'; the broader categories ('a few times a year', 'once or twice a month' etc.) reported in G4 science are less applicable for the secondary school experience where science will be taught in dedicated timetable slots in the same way as mathematics. As with the previous sections, to ease final interpretation of the scales and variables, responses were 'reversed' to present higher scores in line with the claimed direction of association with high achievement scores. For each variable an analysis of frequency of exposure is examined alongside the mean science achievement score. The selections of learning experiences identified through the literature review are clustered under the following headings with associated exploratory data analyses reported in Appendix 4:

Reform-based practice and Discussion

1. Give an explanation about what has been studied
2. Memorizing science facts and principles

Active Learning & Practical Experiments

3. Watch the teacher demonstrate a science experiment or investigation
4. Design or plan a science experiment or investigation
5. Conduct an experiment or investigation

Learning Environment

6. Work problems on our own
7. Work in small groups on an experiment or investigation
8. Listen to the teacher give a lecture-style presentation

Contexts

9. Relating what is learned in science to our daily lives

Technology

10. We use calculators
11. Use a computer for science schoolwork (in and out of school)

Assessment and Feedback

12. Review our homework
13. Have a quiz or test

In the reform movement there is a call for a reduction in 'memorising things' and passive learning, in favour of an increase in active learning, conducting practical experiments and investigations, and ownership of learning experiences. The mean science achievement scores for each variable are analysed to investigate general trends within the data. Some notable findings that can be followed through in the multilevel analyses include:

Explaining

Only a small proportion of G8 students 'never' have opportunities to give explanations about what they are studying. Where there are opportunities for those interactions, the experience is associated with higher student achievement scores in science; highest scores occurring when witnessed 'every or almost every lesson'. This finding appears to support the claim that opportunities to discuss and explain understandings can be beneficial to learners, a feature that can be scrutinised in the multilevel model of science achievement scores. This is a different picture to that reported earlier (G4 science) where memorising appeared to take precedence over explaining in its association with high achievement scores.

Experiments and investigations

Increasing levels of practical work for students, whether watching their teacher, planning the experiment themselves, or conducting experiments, are associated with increasing science achievement scores where the experience occurs in about half the lessons or more. Much as reported for G4 students on this aspect of classroom experience, there is a distinct disadvantage for students who never benefit from such opportunities to engage with scientific practical work; fortunately this only affects small proportions of the cohort but the differences across categories remain statistically significant with 'conducting experiments' providing the greatest explanation of variance:

Watch teacher demonstrate ($F_{3, 58236} = 388.2$, $p < 0.001$).

Plan a science experiment ($F_{3, 57683} = 270.9$, $p < 0.001$).

Conduct experiment or investigation ($F_{3, 57423} = 539.1$, $p < 0.001$).

Working by myself; Working with others

Collaborative group work is a dominant classroom activity with over half the cohort reporting this as happening every, or almost every, lesson. Over 80% of respondents are associated with a higher than average achievement score in science which is quite a contrast to that reported on this variable and its association with mathematics achievement where: (a) much smaller proportions (<25%) indicated engagement at those levels; and (b) the associated achievement scores were lower than in other categories, including the 'never' option that was reported by over a third of the students.

Findings in relation to *lecture-style presentation* and *daily life* for science achievement scores match those presented above within the mathematics achievement data.

Reviewing homework

Evidence of reviewing homework in science appears to offer quiet different patterns and associations with science achievement scores compared to those reported in mathematics classes. Nearly one quarter of the respondents report this experience as a regular feature of their classroom activity, happening on every or almost every lesson, but this is associated with the lowest mean achievement score, which calls into question the value of such experience, or the manner in which it is administered in the science classroom. This is counter to initial observations in mathematics education where reviewing homework is associated with higher than average achievement scores.

4.2.5. ICT experiences (G8)

ICT experiences in school, at home and in public spaces is reported to give a profile of availability and access. Frequency of use specifically for schoolwork, and association with achievement is presented in Table 4.2-5. The distribution of response and associated achievement scores are presented below.

Table 4.2-5: Average of G8 mathematics/science achievement by computer use (home, school, other)

COMPUTER USE			Mathematics		Science	
	N	Percent	Mean	S.D.	Mean	S.D.
PC BOTH AT HOME AND AT SCHOOL	41,300	70.7	497.9	76.2	506.5	76.5
PC AT HOME BUT NOT AT SCHOOL	14,400	24.7	472.5	71.2	478.2	72.7
PC AT SCHOOL BUT NOT AT HOME	1,600	2.8	441.6	71.2	458.0	75.4
PC ONLY AT PLACES OTHER THAN HOME AND SCHOOL	800	1.4	415.3	71.9	431.7	64.8
DO NOT USE PC AT ALL	200	0.4	434.6	63.5	436.7	68.7
Total	58,400	100.0	488.6	76.6	496.8	77.1

As reported for G4 students a significant majority of students make use of computers at both home *and* school, with a further quarter claiming to use computers at home but not in school. Those two categories, covering over 95% of students, are associated with high achievement scores in both disciplines.

The reported frequencies and associated achievement scores in relation to using a calculator or computer are presented in Appendix 4. Over 50 per cent of students report using a calculator in about half their mathematics lessons or more frequently, representing a major shift from that reported at G4 where there was fewer than 10 per cent using this type of technology. Regular use of a calculator appears to be associated with higher mathematics achievement scores. A high proportion of over two thirds of the cohort report 'never' using a computer in mathematics classes, with just over a further quarter using a computer in only some of their lessons. The achievement scores in mathematics for those students who make more regular use of a computer are generally lower than average. A similar pattern of achievement is reported in the science data, calling for inclusion of these *technology* variables in the multilevel analyses to ascertain direction of associations with achievement as frequency of use changes.

Extending this analysis to take account of student's use of computers in support of schoolwork, in and out of school, goes beyond the data reported in Table 4.2-5 and relates patterns of usage across disciplines to achievement scores. The highest achievement scores

in both disciplines are associated with low levels of computer use, from ‘a few times a year’ up to ‘once or twice a month’.

4.2.6. School culture and ethos (G8)

Measures of school culture and ethos are derived along the same lines as reported in 4.1.8; the G8 responses are presented in Appendix 4. Students’ views on their peers putting in their best effort do not appear to be strongly associated with their mathematics or science achievement scores. There appears to be an association between achievement scores and perceived strength of feeling on a positive and supportive culture, where teachers show that they want the best for their learners. Low achievements scores are associated with disagreement (a little, or a lot), although only a small proportion of respondents (~10%) do not feel their teacher wants their children to do their best.

The second measure of school ethos reflects on how ‘safe’ students feel at school, an IEA-derived index-variable of High, Medium and Low perception of being safe as described in 4.1.8. The comparable variables within the G8 data have the ‘B’ prefix. A high perception of safety was assigned on the basis of responding ‘no’ to all five concerns; a low perception of safety assigned if there were three or more concerns over ‘safety’. Sixty per cent of the G8 cohort reports no concerns on safety. Where there are concerns, medium or low, those categories are associated with lower achievement scores, with a more notable association reported within mathematics discipline.

5. Primary Analyses

Science and Mathematics achievement data will be modelled separately, taking subject specific classroom experiences discussed in Chapter 2 into account. Identification of final models will provide insight to effective practices and pedagogies, with associated measures of effect as based on the empirical findings in the survey data. Whilst there is not a direct match of $TIMSS_{2007}$ variables to all of the themes and issues outlined in Chapter 2, there are derived variables and appropriate measures that can be used to evaluate learners' experiences in developing models of students' achievements in science and mathematics.

The EDA of survey items in Chapter 4 identified potential control and explanatory variables for consideration in models of student achievement. These survey and derived variables are explored in multi-level analyses to confirm suitability for inclusion. Thorough diagnostic tests are deployed throughout to ensure the variables do not show threats to validity. Each model of student achievement is developed in stages, starting with the partitioned unconditional model and then adding control and individual predictor variables of interest to ascertain effects on the dependent variable of student achievement. The models are members of the variance components family, where the model splits the total variation into components of variation for each level in the model. There is only one set of random effects (intercepts) for each level in the model; hence the name random intercepts model. A range of control variables are considered and retained as warranted, checking variance partition coefficients at each step, the significance of coefficients for parameter estimates, and confirming subsequent models are an improvement on previous by assessing model comparison using the deviance information criterion (DIC; Spiegelhalter, 2002). As outlined in Chapter 3, the primary method adopted to develop models of achievement uses MCMC methodology; the default mechanism for MCMC method in MLwiN is to use Gibbs sampling for all parameters. This approach permits robust diagnostic assessment of parameter estimates and standard errors, and also provides analyses of accuracy related to chain length. Given the data is normally distributed, estimation of parameters was subsequently computed using maximum likelihood approach to generate models on basis of Iterative Generalised Least Squares (IGLS). The parameter estimates were comparable with those found through MCMC method but considerably more efficient to compute so IGLS was used to develop the final models of Mathematics and Science achievement at G4 and G8. Those final multivariate models consider all the explanatory variables together as a way of arbitrating between competing concepts to draw conclusions from those analyses and measures of effect size.

5.1. Predictor variables

The literature review of substantive issues in learning and teaching of mathematics and science laid claim to particular classroom experiences that were educationally more effective. The survey data does not provide explicit variables for each and every aspect of the literature, but some underlying principles within the theories can be investigated. Taking Watson's delineation between 'traditional' and 'reform' oriented practice (Watson, 2008) as a starting point, the following variables are related to the conceptual framework and provide evidence of support or otherwise for the claims on effective practice. Traditionalists in the 'Math Wars' (Schoenfeld, 2004) placed memorisation and practice as necessary and desirable activities in learning, expressing concern over a proposed shift to process skills at the expense of content coverage. Given 'traditional' experiences have a tendency to be teacher-oriented, placing teacher exposition followed by students working on their own as a typical format, the G4 variables of note are related to '*memorising*' and '*working on my own*', and *'watching the teacher*' where it comes to practical investigations, activities and experiments. At G8 an additional variable on '*lecture-style presentation*' will be added to that dimension of the analysis.

The 'reform' emphasis on project work, with opportunities for group learning activities that included peer-to-peer discussion and discussion between teacher and students, relates to the variables that report on '*working in small groups*' and '*explaining*' concepts or processes. A central argument within the reform movement rests on conceptual understanding, with the claim that active and collaborative learning can lead to that goal. Any associated discussion and explanation will add to students' depth of knowledge and understanding as they develop connected schema and relational understanding over the less-valued instrumental rote learning or memorisation without understanding. When students explain their processes and what they are learning it reveals a lot about their understanding and can of itself enhance understanding, none more so than when pursued within practical experiences (White & Gunstone,1992). Those sentiments link to the variables on '*doing*' or '*conducting*' experiments or investigations in science or other practical activities such as '*measuring*' or '*making tables charts or graphs*' in mathematics. A further role of discussion and explaining falls within feedback on assessment. The contributing variables on that dimension are those that report on frequency of having a '*quiz or test*', and how often students '*review homework*' within their lessons.

A further feature of active and collaborative learning that lies within constructivist theories of learning relates to the social and contextual nature of experience. Students' ability to transfer their knowledge to new applications and contexts is dependent on their depth of

understanding. Views are divided on this issue, over how best to pursue application of knowledge in new contexts, and whether it is of benefit to have a range of experiences *en route* to learning or to consolidate a robust technical understanding as a first step to gaining a genuine understanding (LMS, 1995). This contextual nature of learning can be evaluated through the variable that reports on students relating their learning to '*daily life*'.

Use of ICT in learning and teaching is a final aspect for consideration, to evaluate the claimed benefits of technology in support of learning. The predictor variables that support this aspect are sub-divided into '*calculator*' technology and '*computer*' use in lessons, and '*computers for school work*' where accessed beyond the classroom. There is a further acknowledgment of ICT in support of learning as measured through the control variable on ICT at home.

These 'traditional' and 'reform' practices and experiences can be evaluated through the empirical data to determine whether they provide explanation of variance within reported achievement scores. The conceptual framework in Chapter 2 and exploratory data analysis in Chapter 4 supported key aspects of classroom experience that are now considered in terms of explaining variance in learning and teaching. In summary these are:

Table 5.1-1: Variables considered as predictors

How often do you do these things in your …	
mathematics lessons (G4)	**science lessons (G4)**
Reform-based practice and Discussion	
Explain my answers	Write or give an explanation for something I am studying in science
Memorise how to work problems	Memorising science facts and procedures
Active Learning & Practical Activities	
Measure things in the classroom and around the school	Watch the teacher do a science experiment
Make tables, charts or graphs	Design or plan a science experiment or investigation
	Do a science experiment or investigation
Learning Environment	
Work problems on my own	Work science problems on my own
Work with other students in small groups	Work with other students in a small group on a science experiment or investigation
Technology	
Use a calculator	Use a computer in science lessons
We use computers	Use a computer for science schoolwork (in and out of school)
Use a computer for mathematics schoolwork (in and out of school)	

How often do you do these things in your …	
mathematics lessons (G8)	**science lessons (G8)**
Reform-based practice and Discussion	
Explain our answers	Give explanations about what we are studying
Memorise formulas and procedures	Memorise science facts and principles
Active Learning & Practical Activities	
Interpret data in tables, charts or graphs	Watch the teacher demonstrate an experiment or investigation
Decide on our own procedures for solving complex problems	Design or plan an experiment or investigation
	Conduct an experiment or investigation
Learning Environment	
Work problems on our own	Work problems on our own
Work together in small groups	Work in small groups on an experiment or investigation
Listen to the teacher give a lecture-style presentation	Listen to the teacher give a lecture-style presentation
Contexts (real life)	
Relate what we are learning in mathematics to our daily life	Relate what is learned in science to our daily lives
Technology	
We use calculators	We use computers
We use computers	Use a computer for science schoolwork (in and out of school)
Use a computer for mathematics schoolwork (in and out of school)	
Assessment & Feedback	
Review our homework	Review our homework
Have a quiz or test	Have a quiz or test

Models of science and mathematics achievement scores will be developed with each of those aspects in turn, building on the EDA and fitting the variables as predictors that can explain variance in the achievement scores. Sets of control variables are added as fixed factors to each of the models. These control variables were studied separately and are known to explain variance within models of achievement. The resultant models are developed separately and then reviewed collectively, grouped on the basis indicated above to arbitrate between competing concepts within a theme and to draw conclusions on which aspects of learning and teaching appear to explain variance within the hierarchical models of students' achievement.

5.2. Science Education (G4)

The general multilevel model which allows for school (l), teacher (k), and student (j) effects on achievement, takes 'plausible value' as the lowest level (i) in the model, where every student is assigned five plausible values for each dependent variable in the data set. The form of the general model of science achievement is written as:

$$ASSSCIPV_{ijkl} = \beta_0 + f_{0l} + v_{0kl} + u_{0jkl} + e_{0ijkl}$$

where G4 science achievement is the response variable and β_0 is the overall mean score across all schools. The error in student achievement is presented separately for each level of the model, with f_{0l} being the error attributed to school l, v_{0kl} the error attributed to teacher k (nested within school l), u_{0jkl} a measure of random error at student-level, and e_{ijkl} is a plausible value residual to complete the breakdown of error in the model as a whole. Each of the errors is assumed to be drawn from a Normal distribution with mean 0 and variance σ^2 as shown in the unconditional model for the G4 students in Table 5.2-1.

Table 5.2-1: Unconditional Model of Science Achievement

MCMC Estimation	Unconditional Model A1 (250,000 iterations)		Unconditional Model A2 (350,000 iterations)	
Response: ASSSCIPV	Coefficient	s.e.	Coefficient	s.e.
Fixed Part				
CONSTANT β_0	504.5	2.7	504.5	2.7
Random Part				
Level-4: IDSCHOOL f_{0l}	398.0	141.8	400.5	140.5
Level-3: IDTEACH v_{0kl}	678.0	148.1	675.9	146.2
Level-2: IDSTUD u_{0jkl}	3668.0	98.2	3668.3	98.0
Level-1: IDCASE e_{0ijkl}	880.2	10.7	880.2	10.7
Total Variance	5624.2		5625.0	
DIC:	164942.7		164943.0	
pD:	3217.6		3217.7	

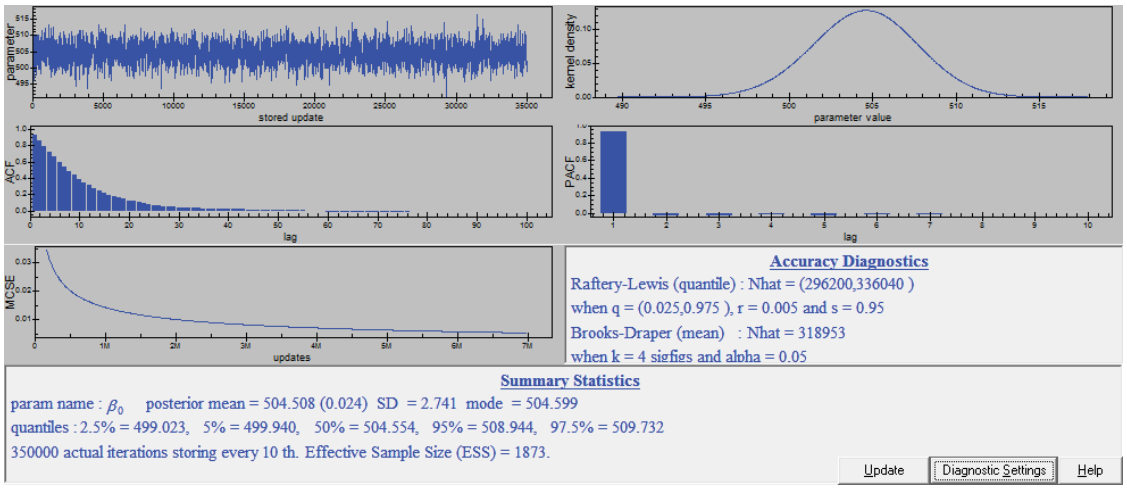
Note: N =16815 cases, 3363 students, with 245 teachers, in 137 schools;

These estimates in Model A_1 and Model A_2 are based on the MCMC statistics following 250,000 and 350,000 iterations respectively. The length of chain was guided by the MCMC diagnostics that detailed desirable number of iterations to secure particular levels of accuracy. Expanded diagnostics for the β_0 parameter are shown in Figure 5.2-1.

The upper left-hand cell reproduces the whole trace for the stored values of the parameter; here it can be seen that the chain is well mixed with the sample space being utilised in full. The goal in any implementation of MCMC method is to take draws from the

joint posterior density once it has converged to stationary. In other words after the 'burn in' period subsequent samples will be safely thought to come from the stationary distribution. The ideal scenario is to have draws from a posterior density that reflects a high degree of 'mixing in a simulation, which is the extent to which a simulated chain traverses the entire parameter space' (Kass, 1998: 94). The upper right-hand cell gives a kernel density estimate of the posterior distribution (equivalent to a smoothed histogram); in this example the density looks to approximately follow a Normal distribution.

Figure 5.2-1: MCMC Diagnostics for β_0 (based on 350,000 iterations in MLwiN)



The second row of boxes plots the autocorrelation (ACF) and partial autocorrelation (PACF) functions. Autocorrelation is a potential problem in MCMC estimation, whereby consecutive values in the chain are not independently distributed, but are rather strongly influenced by preceding estimates. This can be monitored by visual inspection of the trace but more systematically analysed via ACF; here the autocorrelation receded as the chain length increased, showing the parameter estimates as close to being IID data (autocorrelation = 0). The PACF has a spike at lag 1 indicating that Gibbs sampling here behaves like a first order autoregressive time series with a small autocorrelation of less than 0.1.

The third row consists of some accuracy diagnostics. The left-hand box plots the estimated Monte Carlo standard error (MCSE) of the posterior estimate of the mean against the number of iterations. The MCSE is an indication of the accuracy of the mean estimate, where $MCSE = \frac{SD}{\sqrt{n}}$ and n is the number of iterations (350,000). This graph allows the user to calculate how long to run the chain to achieve a mean estimate with a particular desired MCSE. Note the horizontal axis is marked in millions, so the current estimate based on 350,000 iterations, gives a MCSE of around 0.024. The right hand box contains two contrasting accuracy diagnostics. First the Raftery-Lewis analysis (Raftery and Lewis, 1992)

is a diagnostic based on a particular quantile of the distribution. The diagnostic Nhat is used to estimate the length of Markov chain required to estimate a particular quantile to a given accuracy. In MLwiN the default settings for this diagnostic are the 2.5% and 97.5% quantiles that will form a 95% central interval estimate. For this particular parameter the estimated chain length (Nhat) is 296,200 for the lower quantile and 336,040 for the upper quantile. This diagnostic influenced the choice of chain length to secure the accuracy of that central interval estimate. Second there is the Brooks-Draper diagnostic. This diagnostic is based on the mean of the distribution and is used to estimate the length of Markov chain required to produce a mean estimate to 4 significant figures in this case. Here we can see that to quote our estimate as 504.5 (4 sig. fig.) it requires the chain to run for 318,953 iterations; so the diagnostic is satisfied for that parameter. If the settings are changed, to make $k = 3$, the Brooks-Draper Nhat decreases to only 3,190 iterations for the same parameter. In subsequent developments of the model it will be sufficient to work to 3 significant figures; doing otherwise would demand extremely long chains to secure any greater accuracy across a range of parameters. The diagnostic settings (option box on screen shot) can be modified to suit the analysts' needs across all of these diagnostic measures (Kernel density, ACF/ PACF, MCSE, Rafter-Lewis for quantiles, and Brooks-Draper for mean, such as changing the value of *k* for number of significant figures).

The bottom box contains numerical summaries of the data. As well as the posterior mean and its MCSE in parenthesis (confirming the Monte Carlo standard error of 0.024); this box also provides estimates of the mode and median of the posterior distribution. The quantile values of the distribution are given to estimate both 90% and 95% intervals; here the 95% central interval (Bayesian credible interval whose accuracy has already been confirmed through Raftery-Lewis diagnostic) runs from 499.023 to 509.732, working with tolerance of 0.005 as specified by r in the diagnostic settings. The last row of this box includes details of the run length of the Markov chain and an estimate of the effective (independent) sample size (Kass *et al*., 1998). Here the number of stored iterations, every 10^{th} of the 350,000 actual iterations, is divided by a measure of the correlation of the chain to give the effective sample size of 1,873. This big reduction from the actual iterations to the effective sample size reflects the level of autocorrelation and lack of independence in the full chain. In other cases, where the actual iterations and effective sample size are closer in magnitude, the sample of iterations is regarded as equivalent to an independent sample of iterations. The *thinning* process, where only every k^{th} iteration is stored, is used to cut down on the computational memory required when monitoring long runs; in this example it is every 10^{th} iteration but in many of the subsequent models I have used every 5^{th} iteration to provide

fuller accuracy. In terms of accuracy diagnostics, it is worth noting that the parameter mean and standard deviation use *all* the iterated values no matter what thinning factor is used; all other summary statistics and plots in MLwiN are based on the thinned chain only.

The statistics of interest at this stage are the variance components. Two assumptions fundamental to analysis of variance methods are: (a) The residual deviations are normally distributed, and (b) the residual variances are homogeneous. As outlined in Chapter 3.3.1, the analysis of residuals within hierarchical data takes account of the structure and number of cases within each level. Rather than working with the raw residual for an individual, the difference between observed and predicted ($r_{ij} = y_{ij} - \hat{y}_{ij}$), the estimated residual for any chosen level is obtained by multiplying the raw residual by its shrinkage factor, a transformation that gives a shrunken residual. The shrinkage factor k being $\frac{\sigma_{\mu}^2}{\sigma_{\mu}^2+(\sigma_e^2/n_j)}$.

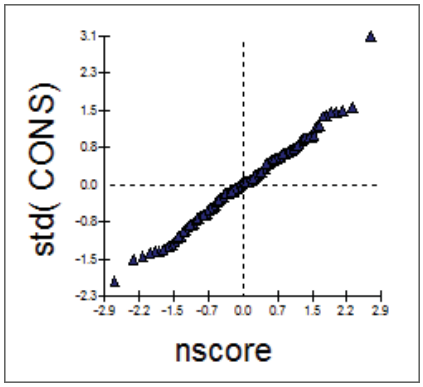
MLwiN calculates the residuals at each level, taking the necessary shrinkage factor into account, and provides options for a variety of graphical displays to inspect them. To check the normality assumption, a normal probability plot of ranked residuals against a normal curve can be inspected. If normality assumption is valid then the points should lie approximately on a straight line. The normal plots for Model A1 are shown in Figure 5.2-2.

Each plot is of the standardised residuals at each level of the hierarchy by normal scores (nscore), where (i) displays plot of residuals at school-level (n=137); (ii) displays comparable plot of residuals at teacher-level (n=245); (iii) plots residuals at student-level (n=3363); and (iv) plots residuals of plausible value as unit-level (n=16815). All four plots satisfy the assumption of normality with only minor deviations in the tails of the higher-level residual plots that are based on a smaller number of units.

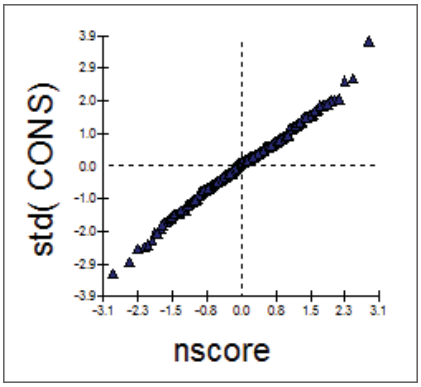
A closer inspection of these plots in tandem with the corresponding plot of standardised residuals against predicted values (fixed part prediction) provides a visual check on homoscedasticity within the model, checking the second assumption that variances are homogeneous. Within those plots in Figure 5.2-3 to Figure 5.2-6, I have highlighted specific schools for discussion, with teachers and pupils in those schools retaining the same colour-coding in their comparable plots.

Figure 5.2-2: Normal Probability plots for unconditional model of science achievement (G4)

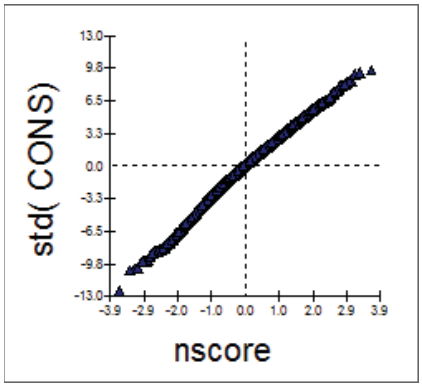
(i) School-level



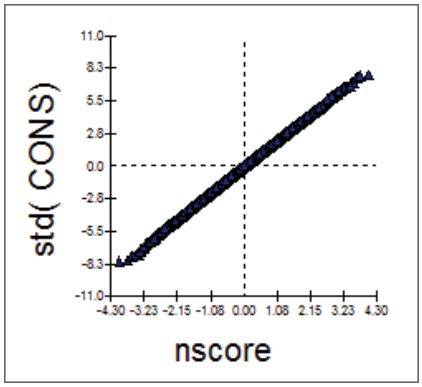
(ii) Teacher-level



(iii) Student-level



(iv) Unit-level (plausible values)



The extreme values in Figure 5.2-3 have been highlighted to make fuller sense of the data and their residuals. First, at the upper end of the plot school ID 147 (BLUE) is highlighted as the highest residual, and school ID 149 (GREEN) as the second highest. At the other end of the ranking school ID 36 (CYAN) is highlighted as the smallest standardised residual and finally school ID 24 (RED) has been highlighted for reasons that will become more apparent when the teacher-level residuals are discussed. The plot of residuals against predicted values in Figure 5.2-4 highlights extreme cases are not necessarily outliers in the model but merely cases that display a high level of variance in relation to others. The school with the highest residual ID 147 (BLUE) happens to be an outlier with one of the average predicted scores in the model of science achievement incorporating response on explain science studied. The main purpose of Figure 5.2-4 is to visually inspect the homogeneity of variances: there is no obvious fanning out of residuals across the range of predicted scores, thus allowing me to conclude that the variances are homoscedastic.

Figure 5.2-3: Normal probability plot of school-level residuals

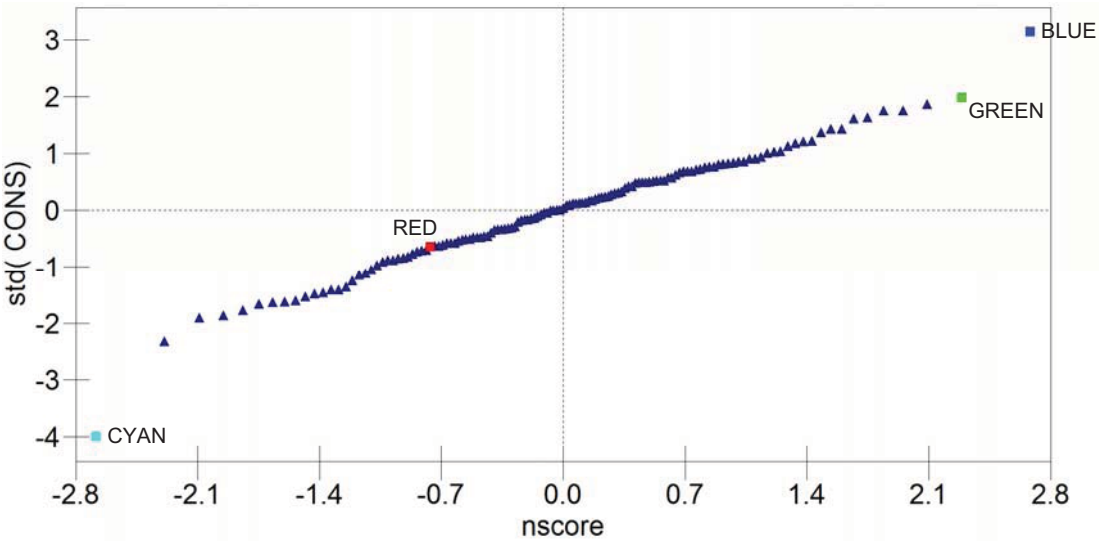
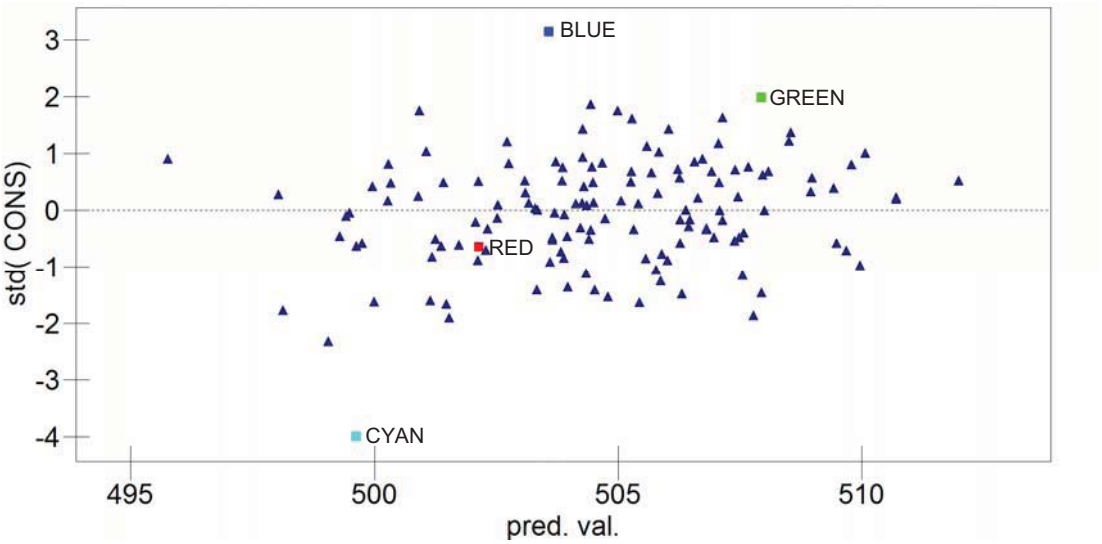


Figure 5.2-4: Standardised residuals (school-level) against predicted value (fixed) for science achievement model (explain science)



The second pair of plots, in Figure 5.2-5 and Figure 5.2-6, provides scope for a similar interpretation. The first comment relates to teachers from school ID 147 (BLUE). There are four teachers from this school, each with a higher than average residual and with two of them ranked 10th and 11th highest out of all teachers in the study. The second comment concerns school ID 24 (RED). One of the five teachers from that school is ranked with the highest variance, yet that does not come through on the school-level plot because the predictions from students taught by the other four teachers reduce the average residuals within that school. The residuals from the other two schools can be interpreted on a similar basis. The second highest ranked school-level residual (ID 149 GREEN) is based on the

residuals from two teachers, each of whom have higher than average residuals. School ID 149 happens to be a high attaining unit as shown in the standardised residual plot against predicted value for science achievement.

Figure 5.2-5: Normal probability plot of teacher-level residuals for science achievement model (explain science)

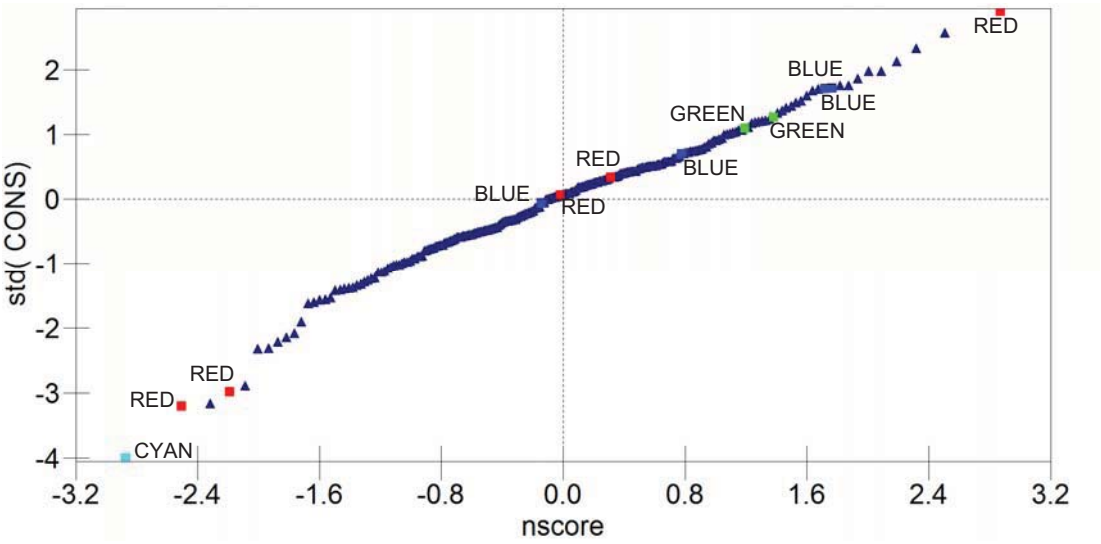
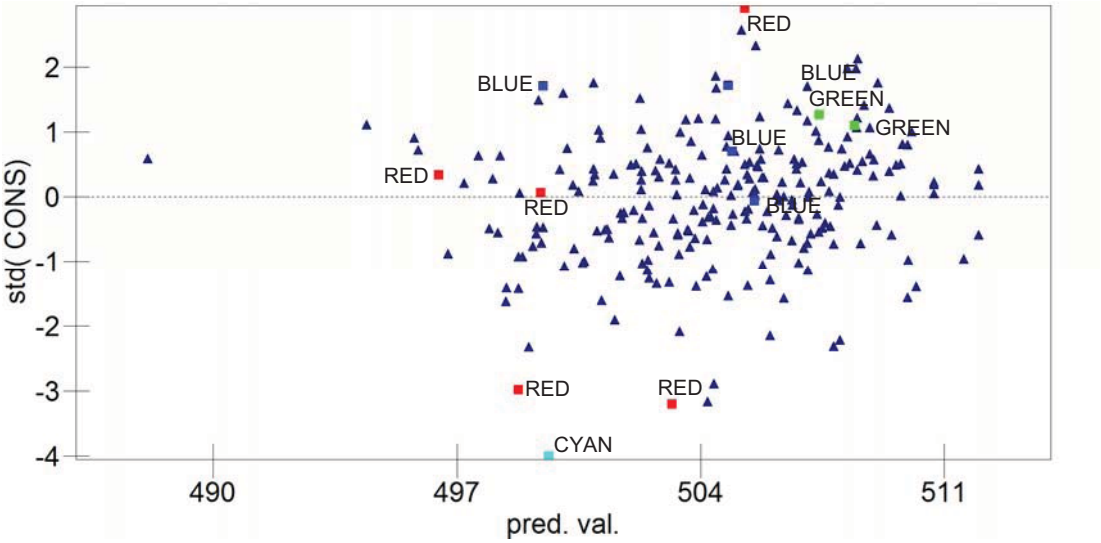


Figure 5.2-6: Standardised residuals (teacher-level) against predicted value (fixed) for science achievement model (explain science)

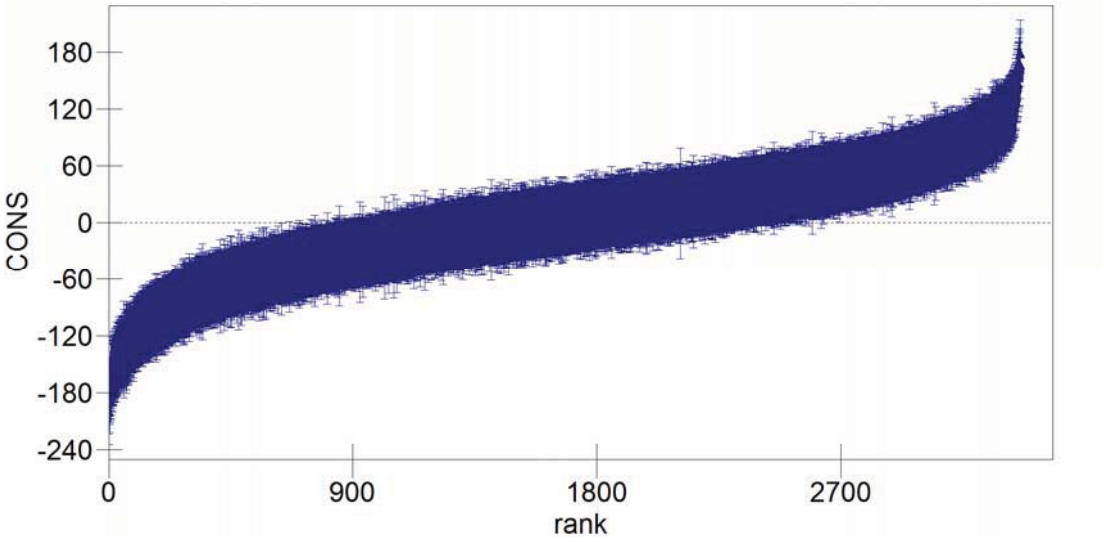


Finally, school ID 36 (CYAN) has only one teacher with the higher-level average residuals based on the predictions from only 13 students. The implication of this small number of contributing predictions and paucity of data is a high shrinkage factor, taking the prediction close to the mean regression value; the fact that this school has both the lowest school- and teacher-level residuals even after shrinkage suggests that its raw performance is

even more extreme. As with the earlier plot, Figure 5.2-6 does not show any severe fanning or clustering of residuals across the range of predicted scores. This visual check confirms that the variances at teacher-level similarly satisfy the assumption that residual variances are homogeneous.

MLwiN provides a variety of graphical displays to inspect residuals. One final display worthy of note plots the residual $+/-1.96$ sd by rank, producing a 'caterpillar plot' as illustrated in Figure 5.2-7.

Figure 5.2-7: Caterpillar Plot of student-level residuals in Model A_1



If the plots do not overlap zero, as witnessed at either ends of the plot for the low and high ranks, those plots represent student departures that are significantly different (at the 5% level) from the overall average predicted by the fixed parameter β_0. In this plot there are around 700 cases with significantly low residuals, and some 730 cases with significantly high residuals. These are the cases that we want to know more about in that they display significantly different patterns of variance in the model.

The proportion of variance at each level of the hierarchical structure gives an indication of where explanations may lie for differences in achievement scores. For example the variance partition coefficient (VPC) at school-level, referring back to Table 5.2-1: Unconditional Model of Science Achievement, is: $\frac{398.0}{5624.2} \cong 0.07$. In the unconditional model A_1, the VPCs show 7% of the variation in student achievement is between schools, 12% is between teachers, 65% between students and the remaining 16% is within-student variation, reflecting the spread in plausible values. Around 19% of the total variance in achievement scores is attributed to differences between schools and teachers.

5.2.1. Explain science studied

Taking each of the science classroom experiences in turn, I start with opportunities to 'Write or give an explanation for something I am studying in science' to evaluate whether frequency of opportunity explains variance in science achievement. Following a run of 250,000 iterations using MCMC method the resultant estimates are set out in Table 5.2-2. For comparison purposes the parameter estimate and partitioned variance from the unconditional model is entered in the same table; renamed as Model A to distinguish from earlier unconditional model as this is computed using a reduced number of cases as dictated by model B.

Table 5.2-2: Comparison between Unconditional Model and opportunities to Explain Science

MCMC Estimation (250,000 iterations)	Unconditional (Model A)		Explain Science (Model B)		
Response: ASSSCIPV	Coefficient	s.e.	Coefficient	s.e.	Sig.
Fixed Part					
CONSTANT β_0	504.1	2.7	488.1	3.7	**
EXPLAIN SCIENCE STUDIED (NEVER)					
A FEW TIMES A YEAR β_1			19.9	3.4	**
ONCE OR TWICE A MONTH β_2			22.4	3.4	**
AT LEAST ONCE A WEEK β_3			11.0	3.6	**
Random Part					
Level-4: IDSCHOOL f_{0l}	392.0	142.5	388.4	133.3	
Level-3: IDTEACH v_{0kl}	673.4	152.5	637.7	140.2	
Level-2: IDSTUD u_{0jkl}	3687.7	95.9	3638.3	94.5	
Level-1: IDCASE e_{0ijkl}	875.2	10.3	875.1	10.4	
Total Variance	5628.3		5539.6		
DIC:	175683.1		175681.9		
pD:	3430.3		3428.6		

Note: N=17920 cases, 3584 students, with 251 teachers, in 137 schools.
** t-value>2.58, sig. level 1%; * t-value>1.96, sig. level 5%

Including the 'explain science studied' predictor variable has reduced the total variance in the model. The percentage reduction in variance in Table 5.2-2 can be computed at each level of Model B. Taking Level-3, the teacher-level variance by way of illustration, the percentage reduction is: $\frac{(673.4-637.7)}{673.4} \cong 0.053$

The reduction of 5% indicates that this proportion of variance is associated with the opportunities to 'explain science' predictor. There is a further reduction in total variance at school-level (1%) and a similar reduction attributed to student-level (1%). Clearly there will be no change in the proportion of variance attributed to the plausible values as this will be

constant across all models. The bulk of variance explained is attributed to teacher-level in the model, reflecting the decision to provide opportunities for students to write or explain aspects of their studies in science primarily rests with their teacher, who is operating within school policy. These variance values, alongside the individual parameter coefficients and standard errors for each response, indicate that opportunities to 'explain science' provide significant explanation of variance in student achievement scores. The categorical variable takes 'never' as its reference category and reports the three separate parameters as providing a significant positive effect on achievement score; t-statistics of 5.8, 6.6 and 3.0 respectively. This analysis confirms the prediction outlined in EDA, with those students who expect to 'explain science' once or twice a month gaining the most in relation to the reference category of 'never' explaining science.

The MCMC diagnostics for these parameter estimates provide confirmation that the desired level of accuracy has been met. The key statistics for accuracy of posterior mean value and Monte Carlo SE come from the Brooks-Draper diagnostic where Nhat = 173,403; the Raftery-Lewis (quantile) statistics have Nhat = (46,275 , 44,470), so this is also clearly satisfied given the chain length of 250,000 iterations. That chain run of 250,000 permits accuracy to 3 sig. fig. making the estimate of 19.9 (3.4) correct within the 95% CI.

The key diagnostics in support of the accuracy for parameter estimates β_1, β_2, and β_3 are:

Parameter	Posterior mean	MCSE	ESS	Raftery-Lewis (quantile) Nhat when q=(0.025,0.975)	Brooks-Draper (mean) Nhat	Sig. Fig.
β_1	19.9	0.021	6,233	(46,275 , 44,470)	173,403	3
β_2	22.4	0.021	5,837	(47,400 , 47,785)	170,355	3
β_3	11.0	0.023	5,820	(51,685 , 47,595)	199,918	3

The diagnostics for each response are inspected in turn and confirm the degree of accuracy is appropriate and valid for the parameter estimates cited in Table 5.2-2. However, a number of control variables are known to have an effect on science achievement scores, so in order to draw substantive conclusions on the overall effect of 'explain science' within the model, those control variables should to be built into the model of science achievement to remove variance that can be attributed to background characteristics.

Student level control variables

There are a number of biographical features that can be included at the student level in the model as well as other derived variables that are intended to serve as proxies for prior attainment and SES. These control variables will be included to strengthen

interpretation of effect attributed to explanatory variables.

The biographical data from the $TIMSS_{2007}$ survey includes gender of student, whether they were born in UK, or came to the country during their early formative years, and a response on how often English is spoken at home. The last of these variables is recorded as 'own language' so that comparisons with other countries can be made in terms of first language at home and also to identify the prevalence and impact of being bilingual where respondents indicated 'almost always' or 'sometimes' in response to the question:

"How often do you speak English <language of test>at home?"

There is not an explicit measure of prior attainment in the studies so although I refer to school- and teacher-effects on the dependent outcome of mathematics (or science) achievement, these terms cannot be used in the same way as cited in much of the literature on school effectiveness research. The interpretation of residuals cannot be put down to the school or attributed to teacher-effects *per se* because the survey data does not take account of intake ability. However, if matched data is made available a 'value-added' analysis can be undertaken, as reported by Sturman *et al.* (2008) in the national report on England's achievement in $TIMSS_{2007}$. Sturman's analysis in that publication included prior attainment data from the National Pupil Database and made the point that their report was a 'value-added' analysis, stressing the point further by noting:

> This means that any reported association between an independent and dependent variable is acting 'over and above' the effect of prior attainment. (Sturman *et al.* 2008: 260)

I certainly acknowledge that incorporating a measure of prior attainment would help to interpret the external effects on student attainment, by accounting for a well-recognised proportion of the observed variance, but nevertheless it would be difficult to lay claim to stronger associations on the basis of such prior measures. The emphasis instead is on trends and associations within the observed data, taking account of survey data to explain variation in attainment between individual students, and between classes, schools and countries as appropriate.

The control variables are entered in batches, with the first wave accounting for 'home' characteristics and a second cluster more closely aligned with individual students and their interest. Each entry was administered individually with usual checks for significance and retention carried out. All subsequent models of G4 science achievement will utilise the same control variables so before progressing I will expand on those control variables as used in this first model presented in Table 5.2-3.

All control variables provide significant explanation and effect on science achievement scores. The first variable takes account of students' place of birth and where

they spent their early formative years. A dichotomous variable (ASDBFORM) takes students who were born in the UK or spent their early formative pre-school years in the UK (≤5 years) as reference category. The model reports a significant negative effect on science achievement for students who arrived in UK aged over 5 years ($\mathrm{t-statistic} = -8.8$). The second variable accounts for spoken language used at home, taking the reference category as 'always' speaking English. All three alternative responses to this categorical variable are significant contributors to the response variable, with a large negative effect where students report 'never' speaking English at home. Students who self-identify themselves as bi-lingual, reporting English as 'almost always' spoken at home, show a positive effect on science achievement.

Responses on the third variable reflect a measure of home facilities and resources. There are four types of resource at home, taken under the headings of 'study tools' (calculator, dictionary, & encyclopaedia), 'ICT' (computer & internet connection), having one's 'own bedroom', and possessing a 'mobile phone'. This first of those is a derived variable that considered the availability of those 'study tools', possessing 'none' through to 'all three', with the model taking 'none' as the reference category. The reference category for ICT was also 'none', i.e. no computer at home, with students reporting possession of a computer with or without internet connection. For the control variables reporting on 'own bedroom' and 'mobile', the reference category of 'yes' is used for both dichotomous responses.

The last 'home' characteristic to be used as a control variable accounts for the number of books the student has at home. This is recognised as a proxy for parental support and for the value placed on education in the home. Given the distribution of responses for this variable, the reference category is taken as the mid-range of '26 to 100 books'. Where fewer books are reported the model indicates a significant negative association with science achievement; also where more than 100 books are reported there was a significant positive association with science achievement. Possessing all three of the study tools is statistically significant at the 5%-level, whereas having a computer with internet, not having their own bedroom, not having a mobile phone, and all responses on the number of books at home below and above the reference group are all statistically significant at the 1%-level.

The second wave of control variables entered in the model is more closely aligned to individuals. There are three broad clusters covering student's perception of safety in school, out-of-school interests and student's gender. The inclusion of gender as a control variable was delayed because there were no discernible differences in science achievement scores on gender alone, it is only after a range of other predictor and control variables have been

included that gender differences are noteworthy. First, students' perception of safety in school is entered as an index measure of high, medium and low; with a high level of safety taken as the reference category. A high perception of safety was assigned on the basis that none of the five potential concerns affected an individual – i.e. concerns over having something stolen; being hit or hurt by another student; forced to do things they didn't want to do by other students; made fun of or called names; or left out of activities by other students. This derived variable is tabled as a proxy for school culture and ethos. A low perception of safety is assigned if the student reports three or more of those concerns. When a low perception of safety is reported, there is a significant negative association with science achievement ($p < 0.01$)

Table 5.2-3: Explain Science (Model B) compared with Model C that includes control variables

MCMC Estimation (250,000 iterations) Response: ASSSCIPV	Explain Science (Model B) Coef.	s.e.	Explain Science +Home +Student (Model C) Coef.	s.e.	Sig.
Fixed Part					
CONSTANT β_0	488.6	3.7	480.2	8.7	**
EXPLAIN SCIENCE STUDIED					
A FEW TIMES A YEAR β_1	19.9	3.4	11.1	3.0	**
ONCE OR TWICE A MONTH	22.4	3.3	13.4	2.9	**
AT LEAST ONCE A WEEK β_3	11.0	3.6	8.7	3.2	**
FORMATIVE YEARS					
ARRIVED IN UK OLDER THAN 5			-42.4	4.8	**
SPEAKS ENGLISH AT HOME					
ALMOST ALWAYS			12.6	2.8	**
SOMETIMES			-18.8	4.1	**
NEVER			-22.8	7.5	**
STUDY TOOLS: CALC, DICT, ENC (NONE)					
ONE OF CALC, DICT, ENCYCLOPAEDIA			4.8	7.7	-
TWO OF CALC, DICT, ENCYCLOPAEDIA			10.2	7.1	-
THREE OF CALC, DICT, ENCYCLOPAEDIA			17.5	7.2	*
ICT AT HOME: COMPUTER, INTERNET(NONE)					
COMPUTER WITHOUT INTERNET			-0.9	5.5	-
COMPUTER WITH INTERNET			24.9	4.8	**
POSSESS OWN BEDROOM (YES)					
NO			-8.2	2.3	**
POSSESS MOBILE PHONE (YES)					
NO			18.6	2.7	**
BOOKS AT HOME (26 TO 100					
NONE OR VERY FEW (0 TO 10			-32.1	3.6	**
ONE SHELF (11 TO 25 BOOKS)			-14.9	2.7	**
TWO BOOKCASES (101 TO 200			19.7	2.7	**
THREE + BOOKCASES (OVER 200 BOOKS)			23.3	2.8	**

MCMC Estimation (250,000 iterations) Response: ASSSCIPV	Explain Science (Model B) Coef.	s.e.	Explain Science +Home +Student (Model C) Coef.	s.e.	Sig.	
PERCEPTION OF SAFETY IN SCHOOL						
MEDIUM			-1.8	2.1	-	
LOW			-17.2	2.7	**	
OUT-OF-SCHOOL INTERESTS -						
(SCLSOCIAL-gm)^1			1.1	2.6	-	
(SCLSOCIAL-gm)^2			-8.6	1.9	**	
(SCLSOCIAL-gm)^3			-4.9	1.8	**	
OUT-OF-SCHOOL INTERESTS - INDIVIDUAL						
(SCLIND-gm)^1			-1.9	1.8	-	
(SCLIND-gm)^2			-11.2	1.8	**	
OUT-OF-SCHOOL INTERESTS – PERSONAL						
(SCLPDEV-gm)^1			1.5	2.2	-	
(SCLPDEV-gm)^2			-13.5	2.1	**	
GENDER (BOY)						
GIRL			-12.3	1.9	**	
Random Part						*Total Proportion Reduction*
Level-4: IDSCHOOL f_{0l}	388.4	133.3	56.7	63.1		0.85
Level-3: IDTEACH v_{0kl}	637.7	140.2	438.8	89.9		0.32
Level-2: IDSTUD u_{0jkl}	3638.3	94.5	2749.0	71.5		0.24
Level-1: IDCASE e_{0ijkl}	875.1	10.4	875.2	10.4		0.00
Total Variance	5539.6		4119.8			0.26
DIC:	175681.9		175632.0	(49.9)		
pD:	3428.6		3380.2	(48.4)		

Note: N=17920 cases, 3584 students, with 251 teachers, in 137 schools;
** t-value>2.58, sig. level 1%; * t-value>1.96, sig. level 5%

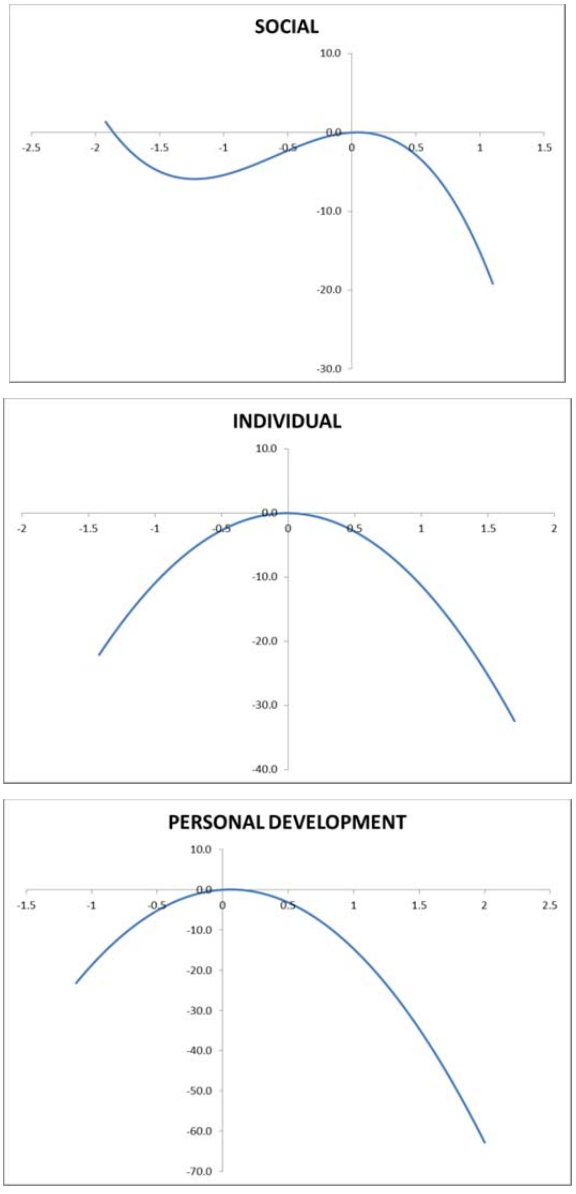
The out-of-school interests are sub-divided into social, individual and personal development groupings, and are modelled on a pseudo-continuous basis as described in Chapter 4. Within the multi-level model it was possible to model on a non-linear basis, and to explore higher powers, retaining as necessary if a significant explanation is provided (based on t-statistic). The social dimension, reflecting levels of social activity through playing or talking with friends and playing sports, is modelled as a cubic function, with the cubic and squared terms of the variable providing significant explanation of variance within the model. The individual and personal development dimensions are modelled as quadratic functions, with the quadratic terms of the variable providing significant explanation of variance. In order to visualise the effect of these standardised functions, a sketch of the combined terms is generated, using the range of values as boundaries for each function. These boundary values are set out in Table 5.2-4 with the sketches in Figure 5.2-8: Out-of-school interests (Social, Individual & Personal Development). These sketches of the

parameter estimates confirm the EDA analysis with extreme values on the scales being associated with lower achievement scores.

Table 5.2-4: Minimum & Maximum of pseudo-continuous variables

Variable	Min	Max	Range
SCLSOCIAL-gm	-1.93	1.10	3.03
SCLIND-gm	-1.43	1.70	3.13
SCLPDEV-gm	-1.12	2.00	3.12

Figure 5.2-8: Out-of-school interests (Social, Individual & Personal Development)



The associated reduction in achievement scores is more pronounced for high

responses, remembering the original reporting scale was linked to time spent on the activities ranging from 'no time' through to '4 or more hours' in any normal school day. Students who spend a moderate amount of time on these out-of-school interests appear to gain most, with the model indicating a significantly negative effect for those spending little or no time on individual interests (watching TV, playing computer games, using internet) or personal development activities (doing jobs at home, reading a book for enjoyment, doing homework). High loadings on those individual and personal development dimensions are reported to be even more detrimental, with a large drop off in achievement scores modelled for personal development; possibly reflecting on individuals who *need* to spend a lot of time on homework and studying outwith their school day.

In contrast, although moderate levels of time playing with friends or playing sport are associated with highest gains on science achievement scores, those students spending little or no time on 'social' activities are not as negatively affected when compared with the other dimensions of out-of-school interests. High loadings on the 'social' variable replicate the large negative effect modelled for 'individual' and 'personal development' variables, although not on the same scale as witnessed on those other dimensions.

The final control variable that has been included in the model of student achievement is a gender parameter, taking boys as the reference category. This is noted as a significant variable in the model with a substantial negative effect reported for girls' science achievement scores in the study. The gender variable provided a significant contribution to the model variance only after other control variables were factored into the model; prior to that action the parameter estimate was not significant.

The model variance is considerably reduced when these control variables are included. The proportion reduction in variance (PVR) when assessing Model C against Model B is: an 85% reduction in school-level variance; a 32% reduction in teacher-level variance; and a 24% reduction in student-level variance. Assessment of the model as a whole is again gauged through inspection of the Deviance Information Criterion (DIC). The measure of fit is significantly improved as the model is developed from Model B to Model C, with a DIC reduction of 49.9 as well as having a further reduction of 48.4 in the model complexity (pD), making Model C a simpler and stronger model of science achievement. There is still unexplained variance in the model, but only 1% is attributed to school-level and 11% remains unaccounted for at teacher-level. The bulk of unexplained variance (67%) continues to be at student-level. Model C has a lower overall variance as a result of taking the predictor and control variables into consideration, but the net effect of this is that the constant variance attributed to the plausible values now makes up a greater proportion of the

unexplained variance, accounting for over a fifth of remaining unexplained variance in the model (21%).

The empirical data shows these control variables to be significant contributors to the variance in the model, a fact that highlights practices as worthy of monitoring, developing or moderating as necessary to strengthen achievement in science studies. The pedagogical practice of getting students to write or give an explanation of topics or concepts studied in science is empirically supported as having a positive impact, with a small but significant effect when students report this practice happening once or twice a month; the effect appears to weaken when the frequency of practice increases to at least once a week, but this still offers a significant benefit when compared with students never having opportunities to explain what they are studying in science.

5.2.2. Combining predictor variables

The analyses from univariate predictor models were computed and then associated statistics from multivariate models using a combination of predictors were analysed. For example, a complex model that requires careful analysis and interpretation of effect is developed when combining 'explain science studied' and 'memorise science facts'. The model that includes both predictor variables and all the home- and student-control variables previously considered shows there is a change in the level of effect from that reported when considered as univariate models. The findings shown in Table 5.2-5 highlight how the effects change when the explanatory variables are combined; as with each of the models the number of cases included in these analyses is determined by the 'complete case' structure, which in this case must satisfy multiple predictors.

The reduction of effect is primarily restricted to the coefficients in the 'explain' predictor, but before proceeding to accept such a model I will provide some analysis of why those effects change in the way they do. When these predictor variables were considered separately, students' opportunity to memorise science facts had the greater effect and this continues to be the case when combined into a single model, with $\Delta \geq 0.35$ for two of the three categories. The same cannot be said for the coefficients associated with categories for explaining science studied: in the combined model, two of those responses no longer offer a significant explanation, with one of them reversed in its direction of effect.

Table 5.2-5: Science combination of 'explain' and 'memorise' (computed coefficients include control variables)

IGLS Estimation	Univariate Models		Multivariate Model EXPLAIN + MEMORISE (Model H)	
Response: ASSSCIPV	Coef.(s.e.)	Sig.	Coef. (s.e.)	Sig.
Fixed Part				
CONSTANT β_0	479.5 (9.0)		469.0 (9.1)	
EXPLAIN SCIENCE STUDIED (NEVER)				
A FEW TIMES A YEAR β_1	11.0 (3.1)	**	7.1 (3.2)	*
ONCE OR TWICE A MONTH β_2	12.2 (3.0)	**	4.2 (3.2)	-
AT LEAST ONCE A WEEK β_3	7.4 (3.3)	*	-2.7 (3.4)	-
Random Part	*Total Var. Reduction*		*Total Var. Reduction*	
Level-4: IDSCHOOL f_{0l}	118.8 (64.21)	0.74	128.0 (65.0)	0.72
Level-3: IDTEACH v_{0kl}	316.4 (72.19)	0.40	317.7 (71.8)	0.39
Level-2: IDSTUD u_{0jkl}	27349.8 (73.75)	0.26	2673.3 (72.0)	0.28
Level-1: IDCASE e_{0ijkl}	875.8 (10.7)	0.0	875.8 (10.7)	0.0
Fixed Part				
CONSTANT β_0	471.8 (9.0)		469.0 (9.1)	
MEMORISE SCIENCE FACTS (NEVER)				
A FEW TIMES A YEAR β_4	10.1 (3.1)	**	8.7 (3.2)	**
ONCE OR TWICE A MONTH β_5	18.8 (3.0)	**	18.3 (3.2)	**
AT LEAST ONCE A WEEK β_6	24.7 (3.0)	**	25.4 (3.2)	**
Random Part	*Variance Reduction*		*Variance Reduction*	
Level-4: IDSCHOOL f_{0l}	128.2 (65.4)	0.72	128.0 (65.0)	0.72
Level-3: IDTEACH v_{0kl}	320.7 (72.3)	0.39	317.7 (71.8)	0.39
Level-2: IDSTUD u_{0jkl}	2684.5 (72.3)	0.27	2673.3 (72.0)	0.28
Level-1: IDCASE e_{0ijkl}	875.8 (10.7)	0.0	875.8 (10.7)	0.0

Note: N=16825 cases, 3365 students, with 250 teachers, in 137 schools;
** t-value>2.58, sig. level 1%; * t-value>1.96, sig. level 5%

To ascertain whether there is a significant interaction between 'memorise' and 'explain' a number of diagnostics were explored. First examining a cross-tabulation of distributions and second, testing an assumption of the hierarchical model that the variables are independent.

The cross-tabulation in Table 5.2-6 is an extension of the EDA carried out in Chapter 4. This highlights a degree of overlap in responses with the largest proportions reported along the common elements on the diagonal.

Table 5.2-6: Cross-tabulation of Memorise and Explain - proportional distribution

		MEMORISE				
EXPLAIN		N	FTY	OTM	ALOW	TOTALS
NEVER (N)	n	1205	710	405	395	2715
	ROW %	*44.4*	*26.2*	*14.9*	*14.5*	*100.0*
A FEW TIMES A YEAR (FTY)	n	805	1525	1300	730	4360
	ROW %	*18.5*	*35.0*	*29.8*	*16.7*	*100.0*
ONCE OR TWICE A MONTH (OTM)	n	460	1190	2180	1935	5765
	ROW %	*8.0*	*20.6*	*37.8*	*33.6*	*100.0*
AT LEAST ONCE A WEEK (ALOW)	n	370	380	1265	1970	3985
	ROW %	*9.3*	*9.5*	*31.7*	*49.4*	*100.0*
	TOTALS	2840	3805	5150	5030	16825
	ROW %	16.9	22.6	30.6	29.9	100.0

Independence can be checked by running the same model for each sub-category of response to determine whether the observed effect is consistent across all categories of the predictor variable. A closer inspection of the distributions and model coefficients within each of the response categories for memorise- and explain-predictor variables is presented in Table 5.2-7 and Table 5.2-8. The reported coefficients are based on the reference category of 'never' (N). There is a broadly similar pattern of effect for each category of EXPLAIN within Table 5.2-7, a pattern that is in line with that shown in Model H that combines both predictor variables with all the home- and student-control variables. There are three entries of note. First, respondents who report a high level of memorising (ALOW) but never explain their science studied are associated with lower effect (21) than reported in combined model (25). Second, students who report low levels of memorising (FTY) *and* low levels of explaining (FTY) appear to be associated with a higher effect (13) than reported in combined model (9). Third, the highest effect in this breakdown is associated with high levels of both memorising and explaining (both ALOW), where the reported model coefficient is 29. This value supports the argument for opportunities to explain their science studied as well as memorising science facts; memorising without such opportunities has a lesser effect in the model of student achievement.

Table 5.2-7: Coefficients for 'Memorise science facts' within 'Explain science studied'

	MEMORISE SCIENCE FACTS		
EXPLAIN SCIENCE STUDIED	FTY (β_1)	OTM (β_2)	ALOW (β_3)
NEVER (N)	8	19	21
A FEW TIMES A YEAR (FTY)	13	16	24
ONCE OR TWICE A MONTH (OTM)	6	20	26
AT LEAST ONCE A WEEK (ALOW)	6	21	29
Model G' (MEMORISE ONLY)	10	19	25
Model H (MEMORISE + EXPLAIN)	**9**	**18**	**25**

The second analysis focuses on explaining science studied as reported in Table 5.2-8. The patterns are less clear in this breakdown, with quite variable effects, although none reported as significantly different from Model H coefficients. An interaction effect was modelled but dropped as it did not add to the explanation afforded by Model H.

Overall, as more explaining is experienced, increasing from a FTY to ALOW, the effect in the model is reduced across all categories of 'memorising' e.g. memorising few times a year (FTY) goes from 15 to 8 to 0. At each level there is a reduced effect attributed to explaining science studied, with lowest impact reported as the level of explaining is at its maximum of ALOW (At Least Once a Week).

Table 5.2-8: Coefficients for 'Explain science studied' within 'Memorise science facts'

	EXPLAIN SCIENCE STUDIED		
MEMORISE SCIENCE FACTS	FTY (β_1)	OTM (β_2)	ALOW (β_3)
NEVER (N)	9	6	-6
A FEW TIMES A YEAR (FTY)	15	8	0
ONCE OR TWICE A MONTH (OTM)	7	7	0
AT LEAST ONCE A WEEK (ALOW)	12	11	5
Model D′ (EXPLAIN ONLY)	11	12	7
Model H (MEMORISE + EXPLAIN)	**7**	**4**	**-3**

Table 5.2-8 provides further support for the argument that memorising *and* explaining has a positive effect when both are witnessed at least once a week (ALOW); reversing the trend that is reported in Model H ($\beta_3 = 5$ *versus -3*). Explaining a few times a year (FTY) has highest effect when in tandem with memorising a FTY or more frequently. Finally, the effect attributed to explaining once or twice a month increases as the level of memorising increases. Collectively these analyses justify the inclusion of both predictors as they provide separate explanation of variance in the model but clearly, in science education, memorising facts is a stronger predictor of science achievement than having opportunities to explain science studied.

This approach of breaking down the analysis is beneficial where there appears to be an association between the predictors. There is no justification for including an interaction term on this occasion, as it has a non-significant effect in the model, but it is important to be able to make sense of changes in the coefficient's values and to be able to interpret the combined effect when building complex hierarchical models. In summary, higher science achievement scores are associated with more frequent opportunity for memorising science facts regardless of the reported level of opportunity to explain science studied; but opportunities to explain, even on a few occasions, appear to further enhance science achievement scores.

5.2.3. Models of G4 Science Achievement

The same processes as outlined in section 5.2.1 were carried out on the G4 science data for each of the previously identified predictor variables. Only complete cases were analysed for each variable, dropping missing cases to maximise the number of respondents included for a complete-case data analysis of that predictor variable and all controls identified as relevant and able to explain variance in the data. The key diagnostics in support of the accuracy of parameter estimates were confirmed at each step of the analysis, extending the MCMC chain to secure the desired level of accuracy to 1 decimal place for each of the predictor variable response coefficients and standard errors. The Brooks-Draper diagnostic was the mainstay of those decisions, varying the number of significant figures accordingly, so that the resultant coefficients were accurate to 1 decimal place.

For each of the parameter estimates reported in Table 5.2-9 the 'effect sizes' and error (95% CI) are presented for the univariate model and multivariate equivalent where variables are considered as a cluster. The effect size for categorical variables is taken as: $\Delta = \frac{\beta}{\sigma}$; and for the standardised pseudo-continuous variables in the model (mean 0 and standard deviation 1), the calculation for effect size is: $\Delta = \frac{2\times\beta}{\sigma}$ (Tymms, 2004). To calculate the effect size for the estimate of β_0, the difference between β_{model} and β_{null} was first computed to quantify the effect before standardising that measure using $\Delta = \frac{\beta}{\sigma}$. For all effect size calculations, in line with discussion in Chapter 3.3.4, the variance at student-level of the unconditional model is used to evaluate Δ i.e. taking $\sigma = \sqrt{u_{0jkl}}$

Table 5.2-9: Summary of Effect Sizes (Models of G4 Science Achievement)

IGLS Estimation		*Univariate Model Coefficients & Significance (Multivariate if change)*			*Univariate Effect size (Multivariate if change)*	
Response: ASSSCIPV		Coef.	s.e.	Sig.	Δ	+/- error
MODEL D: EXPLAIN SCIENCE STUDIED (NEVER)						
A FEW TIMES A YEAR	β_1	10.1 (6.3)	3.1	** (*)	0.17 (0.10)	0.05
ONCE OR TWICE A MONTH	β_2	11.8 (3.8)	3.0 (3.1)	** (-)	0.20 (0.06)	0.05
AT LEAST ONCE A WEEK	β_3	7.1 (-3.1)	3.2 (3.3)	* (-)	0.12 (-0.05)	0.05 (0.06)
MODEL G: MEMORISE SCIENCE FACTS (NEVER)						
A FEW TIMES A YEAR	β_1	9.7 (8.4)	3.1	**	0.16 (0.14)	0.05
ONCE OR TWICE A MONTH	β_2	19.2 (18.9)	3.0 (3.1)	**	0.32 (0.31)	0.05
AT LEAST ONCE A WEEK	β_3	24.9 (25.7)	3.0 (3.1)	**	0.41 (0.43)	0.05

MCMC Estimation		*Univariate Model Coefficients & Significance (Multivariate)*			*Univariate Effect size (Multivariate)*	
Response: ASSSCIPV		Coef.	s.e.	Sig.	Δ	+/- error
MODEL I:WATCH SCIENCE EXPERIMENTS (NEVER)						
A FEW TIMES A YEAR	β_1	20.4 (14.5)	3.2	**	0.34 (0.24)	0.05
ONCE OR TWICE A MONTH	β_2	12.0 (4.7)	3.2 (3.3)	** (-)	0.20 (0.08)	0.05 (0.06)
AT LEAST ONCE A WEEK	β_3	-2.3 (-7.9)	3.3 (3.5)	- (*)	-0.04 (-0.13)	0.05 (0.06)
MODEL J: PLAN SCIENCE EXPERIMENTS (NEVER)						
A FEW TIMES A YEAR	β_1	14.9 (8.5)	2.5 (2.6)	**	0.24 (0.14)	0.04
ONCE OR TWICE A MONTH	β_2	8.7 (3.5)	2.7 (3.0)	** (-)	0.14 (0.06)	0.04 (0.05)
AT LEAST ONCE A WEEK	β_3	1.4 (1.2)	3.2 (3.6)	-	0.02	0.05 (0.06)
MODEL K: DO SCIENCE EXPERIMENTS (NEVER)						
A FEW TIMES A YEAR	β_1	19.3 (13.6)	2.8 (3.0)	**	0.32 (0.22)	0.05
ONCE OR TWICE A MONTH	β_2	20.1 (18.1)	2.9 (3.3)	**	0.33 (0.30)	0.05
AT LEAST ONCE A WEEK	β_3	8.6 (12.4)	3.2 (3.7)	**	0.14 (0.20)	0.05 (0.06)
MODEL N: WORK ON OWN(NEVER)						
A FEW TIMES A YEAR	β_1	14.8 (12.9)	3.9	**	0.24 (0.21)	0.06
ONCE OR TWICE A MONTH	β_2	17.7 (16.2)	3.6	**	0.29 (0.27)	0.06
AT LEAST ONCE A WEEK	β_3	18.6 (18.2)	3.5	**	0.31 (0.30)	0.06
MODEL O: WORK IN GROUP (NEVER)						
A FEW TIMES A YEAR	β_1	12.2 (11.6)	3.6	**	0.20 (0.19)	0.06
ONCE OR TWICE A MONTH	β_2	11.5 (9.8)	3.5 (3.6)	**	0.19 (0.16)	0.06
AT LEAST ONCE A WEEK	β_3	1.7 (0.4)	3.6 (3.6)	-	0.03 (0.01)	0.06
MODEL P: COMPUTER USE (NEVER)						
A FEW TIMES A YEAR	β_1	5.0	2.5	*	0.08	0.04
ONCE OR TWICE A MONTH	β_2	-4.0	2.8	-	-0.07	0.05
AT LEAST ONCE A WEEK	β_3	-19.9	3.6	**	-0.33	0.06

** t-value>2.58, sig. level 1%; * t-value>1.96, sig. level 5%

Reform-based practice and discussion

Memorising facts in science offers the greatest explanation of variance in the multivariate models with effect size $\Delta = 0.43$. The z-score percentiles can be referenced to aid interpretation of the reported effect size, noting that the effect score gives the number of standard deviations above the average student in the control group. For example 'Memorise Science Facts (at least once a week)' has an effect size of 0.43 that translates into 0.43 standard deviations above the average student in the control group i.e. a z-score of 0.50 + 0.43 = 0.93. This gives the percentile value of $P(Z \leq 0.93) = 0.66$ from the standard

Normal Distribution function; the predictor response exceeds the scores of 66% of the control group. Memorising facts in science is very clearly a significant activity with all responses providing significant explanation and substantial effect as frequency of opportunity increases. The joint exposure to memorising and explaining the science studied can have an increased effect in the model. For example explaining a 'few times a year' in conjunction with memorising appears to offer a larger overall effect in the model ($\Delta = 0.10$).

Active learning and practical activities

Evaluating effects for the cluster of variables linked to science experiments or investigations (watch, do, and plan), it is evident that the most consistent impact comes through students doing science experiments. Effect sizes range from $\Delta = 0.20$ to $\Delta = 0.30$ for doing science experiments within the multivariate model. Opportunities to watch their teacher demonstrate an activity a few times a year is associated with a strong effect in the multivariate model, where $\Delta = 0.24$ is equivalent to having the predictor response exceed the scores of 59% of the reference group ($P(Z \leq 0.74) = 0.59$). However, as the frequency of opportunity to watch demonstrations increases, the strength of association with achievement reduces quite rapidly. A similar pattern of association is noted for planning experiments and investigations; a positive and significant effect when some opportunities are afforded ($\Delta = 0.14$) but diminishing benefit when witnessed as a regular feature of learning.

Learning environment

In terms of classroom organisation, all three student responses to 'working on their own' in a science setting provided significant explanation in the multivariate model with effect sizes in the range of $\Delta = 0.21$ to 0.30. Opportunities for group work appear to offer less explanation but still provide a significant contribution when experienced up to once or twice a month; beyond that frequency of exposure the effect is negligible ($\Delta = 0.01$).

Technology

The role of ICT (univariate model of computer use) in learning and teaching does not appear to have any positive association with achievement in G4 Science, with only a marginal effect when witnessed a few times a year and a significant negative association when computers are used regularly. The absolute value of the effect is one of the largest in any of the univariate models of science achievement, $\Delta = -0.33$ equivalent to having 63% of the reference group (those who never use computers in class) outperforming students who claim to make use of computers at least once a week.

5.2.4. Relate science to real life

The last variable for potential inclusion in a model of science achievement concerns the frequency of lessons that students' teachers relate science lessons to daily life. Taking 'some lessons' as reference category the accuracy diagnostics for the two predictor parameters are presented in Table 5.2-10. Clearly the chain length of 300,000 iterations falls well short of the Brooks-Draper (Nhat) requirement to report parameters correct to 2 significant figures but the Raftery-Lewis boundaries are satisfied for the reported posterior mean value. Note the low effective sample size (ESS) for both estimates highlights the auto-correlation and poor mixing in the chain, calling for considerably longer runs should this variable be included in any model.

Table 5.2-10: MCMC Diagnostics for 'REAL LIFE'

Relate to real life (science)				N=15,085 cases; 300,000 iterations		
Parameter	Posterior mean	MCSE	ESS	Raftery-Lewis (quantile) Nhat when q=(0.025,0.975)	Brooks-Draper (mean)	
					Nhat	Sig. Fig.
β_1	1.4	0.074	2,001	(115,695 , 121,655)	2,559,623	2
β_2	5.5	0.084	1,775	(131,120 , 125,595)	3,225,648	2

As noted in the EDA of Chapter 4 there is a very equitable distribution across the three categories with minimal variance in mean achievement scores. Including this variable in the model does not improve the DIC and the model complexity is marginally higher having added these two categorical variables. The parameter estimates calculated through MCMC method are not significant and so this variable is dropped from consideration.

5.2.5. Modelling all predictor and control variables (G4 science)

In order to evaluate competing themes within the data and to identify the principal associations with science achievement, all background control variables and previously examined predictor variables are modelled together. This is a natural extension of the clusters that were analysed as groups in 5.2.3, i.e. the multivariate models presented in Table 5.2-9.

The findings are presented in full in Appendix 5.1. In summary, the findings from this model are presented in Table 5.2-11 where significantly positive and negative associations with science achievement are presented against predictor responses to highlight substantive features of learning and teaching experiences.

Table 5.2-11: Summary of key findings in model of G4 science achievement data with all control and predictor variables included

Response	Effect size	
	Negative association	Positive association
Memorise science facts(never)		
Once or twice a month		0.22
At least once a week		0.41
Watch science experiments (never)		
A few times a year		0.24
At least once a week	-0.13	
Plan science experiments (never)		
A few times a year		0.10
Do science experiments (never)		
A few times a year		0.21
Once or twice a month		0.26
At least once a week		0.18
Work on own (never)		
A few times a year		0.15
Once or twice a month		0.21
At least once a week		0.25
Computer in science lesson (never)		
About half the lessons	-0.14	
Every or almost every lesson	-0.33	

In primary science education it is apparent that the memorisation of science facts is strongly associated with achievement, with the reported effect being equivalent to exceeding the scores of between 59 and 66 per cent of the control group. Over sixty per cent of G4 students identified memorisation as a feature of learning at least once or twice a month i.e. within the two categories reported above.

Another feature of students' experience that has retained significance when all predictors are combined, concerns the handling of practical experiments and investigations. Opportunities for students to do science experiments and investigations are strongly associated with achievement, offering the highest effect sizes within the cluster on practical activities; all three response categories are significant contributors to the model of science achievement. Opportunities to watch their teacher demonstrate an activity a few times a year is also associated with a strong effect in the combined model, retaining the effect size of $\Delta = 0.24$ reported in section 6.2.3 for the multivariate cluster on active learning and practical activities. However, when the frequency of 'watching a teacher do experiments' increases to at least once a week there is a negative association with science achievement; an experience that affects almost thirty per cent of the cohort.

On learning environment, where students report frequency of working on own or in groups, only the former offers a significant association with science achievement in the combined model. The data in Appendix 4.3 (Tables 4.3-9 and 4.3-10) note students having a comparable exposure to working on own and working in groups, and indicate an association with higher achievement scores from both type of classroom organisation. However in the combined model of G4 science achievement we see that working in groups does not offer any significant association with achievement, whereas opportunities to work science problems on own is strongly associated with higher achievement scores across all three response categories; effect size of $\Delta = 0.15$ rising to $\Delta = 0.25$ for working on own at least once a week.

The role of ICT in support of science school work presents an increasingly negative association with achievement as the frequency of use rises to every or almost every lesson. Although that category only affects fewer than ten per cent of students, increasing to just over a quarter of students when considering both categories, what is of greater interest is the finding that computer use in class is not significantly associated with any increase in science achievement scores.

5.3. Mathematics Education (G4)

The general multilevel model for mathematics achievement has an identical structure to that described in 5.2, with 'plausible value' taken as the lowest level (*i*) in the model, nested within student (*j*), teacher (*k*), and school (*l*) levels of analysis. The plausible value for the G4 mathematics data is ASMMATPV, giving the general model of mathematics achievement as:

$$ASMMATPV_{ijkl} = \beta_0 + f_{0l} + v_{0kl} + u_{0jkl} + e_{0ijkl}$$

where β_0 is the overall mean score across all schools and the error in student achievement is presented as in section 5.2. The unconditional model of mathematics achievement for G4 students is shown in Table 5.3-1 .

Table 5.3-1: Unconditional Model of Mathematics Achievement

Estimation		Unconditional Model A_1 (IGLS)		Unconditional Model A_2 (MCMC 350,000 iterations)	
Response: ASMMATPV		Coefficient	s.e.	Coefficient	s.e.
Fixed Part					
CONSTANT	β_0	498.1	2.8	498.2	2.8
Random Part					
Level-4: IDSCHOOL	f_{0l}	424.7	143.4	416.0	142.1
Level-3: IDTEACH	v_{0kl}	714.7	135.7	746.9	151.2
Level-2: IDSTUD	u_{0jkl}	4146.0	104.0	4149.0	104.3
Level-1: IDCASE	e_{0ijkl}	706.5	8.3	706.6	8.3
Total Variance		5991.9		6018.4	
-2*LOGLIKELIHOOD		184240.6			
DIC:				174959.4	
pD:				3534.5	

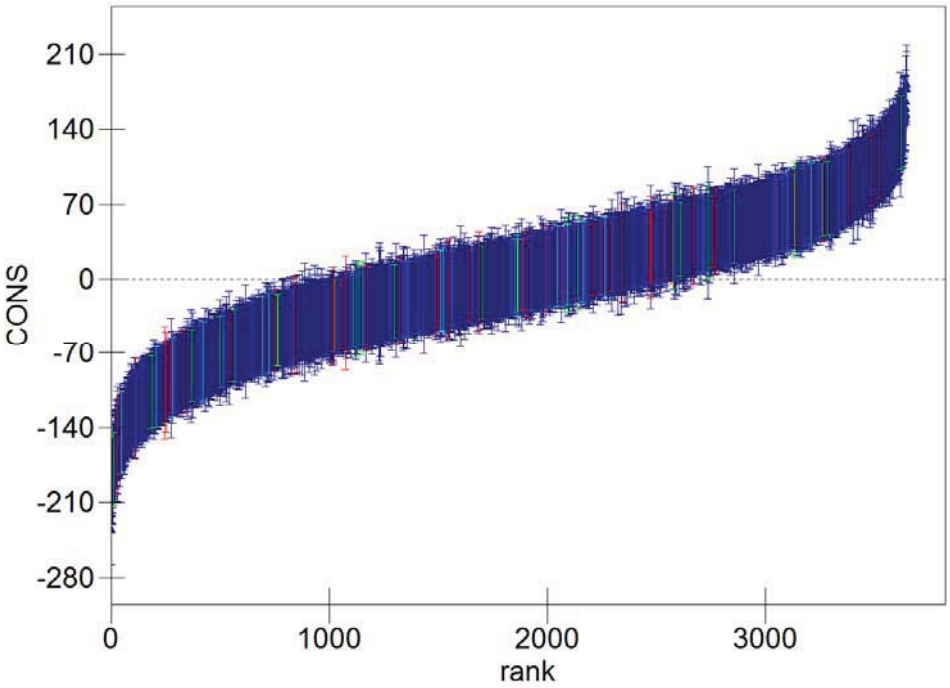
Note: N =18240 cases, 3348 students, with 256 teachers, in 137 schools;

These estimates are computed using different techniques, with Model A_1 generated via Iterative Generalised Least Squares method (IGLS) and Model A_2 derived from a Monte Carlo Markov Chain method (MCMC) as used throughout previous section. As far as the point-estimates are concerned there is little difference between the methods because the data distribution is close to normal, which makes the IGLS methodology fairly robust and considerably more efficient given the MCMC run time for a chain length of 350,000 iterations. The differences occur however in the analysis of variance in the hierarchical structure of the model, with the MCMC methodology providing a more accurate breakdown

of the variance components. The MCMC diagnostic checks provide support for the accuracy of estimates and, as in section 5.1, these will also be used to confirm assumptions on variance components, checking the residual deviations are normally distributed and that the residual variances are homogeneous.

As with the analysis of G4 Science data, Figure 5.3-1 presents a 'caterpillar plot' of the residuals +/- 1.96 s.d. by rank. The plots that do not overlap with zero are deemed to have residuals that are significantly different (at the 5% level) from the overall average predicted score, the fixed parameter in model. In this particular plot there are around 870 cases with significantly low residuals, and around 670 cases with significantly high residuals. This represents a different picture from that presented in Figure 5.2-7: Caterpillar Plot of student-level residuals in Model A_1 (Science) in two ways. First, the predicted average score for mathematics (498.2) is markedly lower than the predicted score for science achievement (504.5), but more notably, lower than the TIMSS international average of 500. Second, the distribution of residuals in the mathematics data shows a higher number cases are significantly different from the predicted average than witnessed in the science achievement data where the reported tails comprised of around 700 cases; the ~870 cases identified in the lower tail of the mathematics data account for just over 25% of the student-level data so there is clearly scope to identify what lies behind those deviations from the reported average.

Figure 5.3-1: Caterpillar Plot of student-level residuals in null Model A_2 (Mathematics)



Referring back to Table 5.3-1: Unconditional Model of Mathematics Achievement,

the variance partition coefficients (VPCs) can be computed. In the unconditional model A_2, the VPCs show 7% of the variation in student achievement is between schools, 12% is attributed to teacher-level variance, 69% is between students and the remaining 12% within student cases, as before, reflects the variation within the five plausible values assigned to each student. As an analyst, it is those proportions of between-school, between-teacher, and between-student variances that need to be targeted to identify plausible explanations for differences in achievement scores.

5.3.1. Models of G4 Mathematics Achievement

Models of mathematics achievement are developed from the Unconditional Model of Mathematics Achievement in Table 5.3-1. The same set of control variables as used in section 5.1 are added as fixed factors in each of the models. The resultant models of mathematics achievement scores in Table 5.3-2 present the univariate and multivariate cluster statistics as an interim guide to associations with mathematics achievement. These data indicate which aspects of learning and teaching and classroom experience appear to explain variance within the hierarchical models of G4 mathematics achievement.

Table 5.3-2: Summary of Effect Sizes (Models of G4 Mathematics Achievement)

MCMC Estimation	*Univariate Model Coefficients & Significance (Multivariate where changes occur)*			*Univariate Effect size (Multivariate where changes occur)*	
Response: ASMMATPV	Coef.	s.e.	Sig.	Δ	+/- error
MODEL B: EXPLAIN ANSWERS (NEVER)					
SOME LESSONS	13.6 (12.6)	3.4 (3.5)	**	0.21 (0.20)	0.05
ABOUT HALF THE LESSONS	17.1 (14.9)	3.7	**	0.27 (0.23)	0.06
EVERY OR ALMOST EVERY LESSON	10.6 (6.8)	3.8 (3.9)	** (-)	0.17 (0.11)	0.06
MODEL C: MEMORISE PROCEDURES ETC. (NEVER)					
SOME LESSONS	13.6 (11.2)	4.5 (4.5)	** (*)	0.21 (0.18)	0.07
ABOUT HALF THE LESSONS	14.1 (11.5)	4.5 (4.6)	** (*)	0.22 (0.18)	0.07
EVERY OR ALMOST EVERY LESSON	23.6 (21.7)	4.4 (4.6)	**	0.37 (0.34)	0.07
MODEL E:MEASURING (NEVER)					
SOME LESSONS	9.1 (6.0)	3.1	** (-)	0.14 (0.09)	0.05
ABOUT HALF THE LESSONS	-13.1 (-13.7)	4.2 (4.3)	**	-0.21 (-0.22)	0.07
EVERY OR ALMOST EVERY LESSON	-30.2 (-28.4)	4.7 (4.8)	**	-0.48 (-0.45)	0.08 (0.09)
MODEL F: TABLES, CHARTS & GRAPHS (NEVER)					
SOME LESSONS	35.6 (32.7)	4.1	**	0.56 (0.52)	0.07
ABOUT HALF THE LESSONS	29.2 (28.9)	4.4	**	0.46	0.07
EVERY OR ALMOST EVERY LESSON	12.4 (20.5)	4.8 (4.9)	**	0.20 (0.32)	0.08

MCMC Estimation	*Univariate Model Coefficients & Significance (Multivariate where changes occur)*			*Univariate Effect size (Multivariate where changes occur)*	
Response: ASMMATPV	Coef.	s.e.	Sig.	Δ	+/- error
MODEL N: WORK ON OWN(NEVER)					
SOME LESSONS	36.7 (31.2)	7.9	**	0.58 (0.49)	0.13 (0.12)
ABOUT HALF THE LESSONS	41.6 (36.5)	7.9 (7.8)	**	0.66 (0.58)	0.13 (0.12)
EVERY OR ALMOST EVERY LESSON	53.3 (48.7)	7.7	**	0.84 (0.77)	0.12
MODEL O: WORK IN GROUP (NEVER)					
SOME LESSONS	13.7 (16.9)	4.3	**	0.22 (0.27)	0.07
ABOUT HALF THE LESSONS	14.8 (18.3)	4.6	**	0.23 (0.29)	0.07
EVERY OR ALMOST EVERY LESSON	-7.7 (-3.3)	4.6	-	-0.12 (-0.05)	0.07
MODEL K: USE A CALCULATOR IN CLASS (NEVER)					
SOME LESSONS	13.8 (14.0)	2.4	**	0.22	0.04
ABOUT HALF THE LESSONS	-1.4 (2.5)	4.7	-	-0.02 (0.04)	0.07
EVERY OR ALMOST EVERY LESSON	-44.3 (-29.2)	7.6 (7.7)	**	-0.70 (-0.46)	0.12
MODEL P: USE A COMPUTER IN CLASS (NEVER)					
SOME LESSONS	-0.8 (-2.5)	2.4	-	-0.01 (-0.04)	0.04
ABOUT HALF THE LESSONS	-14.7 (-14.9)	3.8	**	-0.23	0.06
EVERY OR ALMOST EVERY LESSON	-40.4 (-35.1)	4.4 (4.5)	**	-0.64 (-0.55)	0.07

** t-value>2.58, sig. level 1%; * t-value>1.96, sig. level 5%; - t-value is not significant

Learning environment

The most significant explanatory factor is the student response to 'working on problems on my own'. All three responses provide a significant contribution with effect sizes in the multivariate model ranging from $\Delta = 0.49$ to 0.77. On classroom organisation, it is noted that students' response to working in small groups also provides significant explanation of variance in the model of G4 mathematics achievement. Two of the responses are associated with significant coefficients in relation to 'never' having such an opportunity, with largest effect size ($\Delta = 0.29$) attributed to this occurring in 'about half the lessons', reducing to $\Delta = 0.27$ when only in 'some lessons'. There is marginal negative association when students work in small groups 'every or almost every lesson', but this is a non-significant finding. Overall, this represents a similar pattern of response as witnessed in the G4 science model, but in mathematics there is a stronger weight of evidence in support of group work.

Active learning and practical activities

Another significant contributor to the model of mathematics achievement relates to presentation of data where students acknowledge opportunities to make tables, charts or graphs in their study of information handling or statistics. Each of the responses to using data collected, and having ownership of the development and presentation of data, provides significant explanation of variance in the multivariate model of the G4 mathematics data with effect sizes of $\Delta = 0.52$, 0.46 and 0.32 for 'some lessons', 'about half the lessons', and 'every or almost every lesson' respectively. The other practical component, measuring things in the classroom, is negatively associated with mathematics achievement; this type of experience does not appear to have the same degree of transferability that data presentation afforded in the model.

Reform-based practice and discussion

The next grouping of significant explanation of variance in the ranked order of effect size concerns 'memorising' and 'explaining'. Students memorising formulae, processes and how to work problems (procedures) is the more dominant of the two variables, but having students explain answers in 'about half the lessons' has an effect size of $\Delta = 0.23$, which equates to those students exceeding the scores of 59% of the control group. This distinguishes the results from those reported in the science models where explaining science studied had less of an effect – explaining appears to be important in the mathematics model, although when frequency increases to every lesson the significance of that activity is reduced ($\Delta = 0.11$). This latter pattern of response is perhaps a signifier of general weakness in that the student is constantly being asked to explain their answers, or at least that may be the student's perception of the situation.

Technology

The role of ICT in learning and teaching mathematics has a limited effect, similar to that reported for G4 science. Using a computer does not offer any positive association with achievement in G4 mathematics, with all three responses indicating a negative association with achievement. The absolute value of the effect is equivalent to having 71% of the reference group outperforming students who regularly make use of computers in their mathematics class. Opportunities to use a calculator are associated with some benefit, in that there is a positive association with mathematics achievement when calculators are used in 'some lessons'. When used in half the lessons or more the association with achievement changes, presenting a similar picture to that witnessed with use of computers, namely the

multivariate ICT model presents a strong negative association with mathematics achievement where calculators are used regularly in mathematics classes (every or almost every lesson: $\Delta = -0.46$).

Relate learning in mathematics to daily life

Much as reported within the science data, inclusion of this teacher-level variable does not offer any significant explanation of variance or improve the model fit to the empirical data. There is a very equitable split across the response categories with minimal variance in mean achievement scores reported in EDA. Taking 'some lessons' as the reference category the parameter estimates calculated through MCMC method are not significant:

'About half the lessons' $\beta_1 = -7.6$ $(s.e. = 6.3)$

'Every of almost every lesson' $\beta_2 = 0.78$ $(s.e. = 6.9)$

This variable is consequently dropped from further consideration as an explanatory for variance within models.

5.3.2. Modelling all predictor and control variables (G4 mathematics)

The principal associations with mathematics achievement can be identified by examining the contributions made by all the control variables and predictor variables when modelled together. This is a natural extension of the clusters that were analysed as groups in Table 5.3-2: Summary of Effect Sizes (Models of G4 Mathematics Achievement).

The findings from this full model are presented in Appendix 5.2. In summary, the findings from this model are presented in Table 5.3-3 where positive and negative association with mathematics achievement are presented against predictor responses.

Table 5.3-3: Summary of key findings in model of G4 mathematics achievement data with all control and predictor variables included

Response	Effect size	
	Negative association	Positive association
Explain answers (never)		
Some lessons		0.15
About half the lessons		0.22
Every or almost every lesson		0.17
Memorise procedures etc.(never)		
Every or almost every lesson		0.30
Measure things in class (never)		
About half the lessons	-0.26	
Every or almost every lesson	-0.44	
Make tables, charts & graphs (never)		
Some lessons		0.39
About half the lessons		0.32
Every or almost every lesson		0.22
Work on own (never)		
Some lessons		0.39
About half the lessons		0.47
Every or almost every lesson		0.66
Work in group (never)		
Some lessons		0.17
About half the lessons		0.23
Use a calculator in class (never)		
Some lessons		0.20
Every or almost every lesson	-0.34	
Computer in class (none)		
About half the lessons	-0.20	
Every or almost every lesson	-0.47	

In primary mathematics education the profile of association with achievement in terms of reform-based practice presents a different picture from that reported in G4 science.

Although memorising formulae and procedures is significantly associated with mathematics achievement, with $\Delta = 0.30$ for memorising every or almost every lesson, it is noteworthy that all three response categories for explaining answers in class are significantly associated with raised achievement. Only an eighth of students report 'never' having to explain their answers but almost a quarter of the cohort are in the category associated with a stronger effect in the model, when explaining is witnessed in about half the lessons ($\Delta = 0.22$).

Opportunities for students to engage in active learning and practical activities through making tables, charts and graphs are positively associated with achievement. The other contributing variable in that cluster, measuring things in class, is negatively associated with achievement when students report that as a regular experience, occurring in half the lessons or more frequently; however only small numbers of students fall into those categories with less than ten per cent in each. The contribution that making tables, charts and graphs offers in support of data handling and statistics, and the application of mathematical skills, is worthy of wider consideration given those high effect sizes ($\Delta = 0.22\ to\ 0.39$) that link the experience to high mathematics achievement.

On learning environment, where students report frequency of working on own or in groups, both approaches provide significant association with mathematics achievement in the combined model. In this combined model of G4 mathematics achievement we see that working problems on my own is the more significant predictor with effect sizes of $\Delta = 0.39\ to\ 0.66$, the upper value being equivalent to exceeding the scores of 75% of the control group. A fairly balanced approach is witnessed in G4 mathematics education, with a significant positive association attributed to working in groups as well as having opportunities to work on own. The two significant categories of 'some lessons' and 'about half the lessons' whose scores are positively associated with working in groups account for just fewer than seventy per cent of the cohort. This is in contrast to that reported in 5.2.3 where a small negative association with achievement scores was noted for opportunities to work in groups.

The role of ICT in support of mathematics class work presents a negative association with achievement where use of computers in class is reported. However, a positive association with achievement is noted for calculator use in some lessons, an expereince around sixty per cent of the cohort report as a feature of their learning; more frequent use appears to be detrimental to achievement scores with a strong negative association noted ($\Delta = -0.34$) where calculator use extends to every or almost every lesson and students may become dependent on that technological tool.

5.4. Science Education (G8)

The general multilevel model of G8 Science achievement follows a similar structure to that outlined in earlier sections that allows for school (*l*), teacher (*k*), and student (*j*) effects on achievement. In this primary analysis the case-level measure BSSSCIPV is the 'plausible value' that is taken as the lowest level (*i*) in the model. The form of the general model of science achievement is therefore written as:

$$BSSSCIPV_{ijkl} = \beta_0 + f_{0l} + v_{0kl} + u_{0jkl} + e_{0ijkl}$$

where β_0 is the overall mean score across all schools and the random error in student achievement is presented separately for each level of the model, with each of the errors assumed to be drawn from a Normal distribution with mean 0 and variance σ^2 as shown for the unconditional model for the G8 students in Table 5.4-1.

These estimates in Model A_1 and Model A_2 are based on the IGLS and MCMC estimation methods and report very comparable estimates and breakdown of variance across the four nested levels of the model. Around half of the total variance is attributed to student-level of the model, with a further third at school-level. This presents a quite different partitioning of variance to that witnessed in the G4 data for science and indeed to the G8 mathematics data discussed in section 6.5, where the bulk of the variance in the model is at teacher-level; here we have less than 10% of the total variance at teacher-level of the model whereas about a third of total variance is between schools, highlighting the potential to identify school differences that might explain a large proportion of variance in the model. Another significant difference between G4 and G8 data concerns the number of teachers, with 777 teachers included in the models for G8 Science Achievement compared to only 251 in the comparable models for G4 Science Achievement; this reflects an increase in specialist teachers in the secondary sector relative to primary setting.

Table 5.4-1: Unconditional Model of GR8 Science Achievement

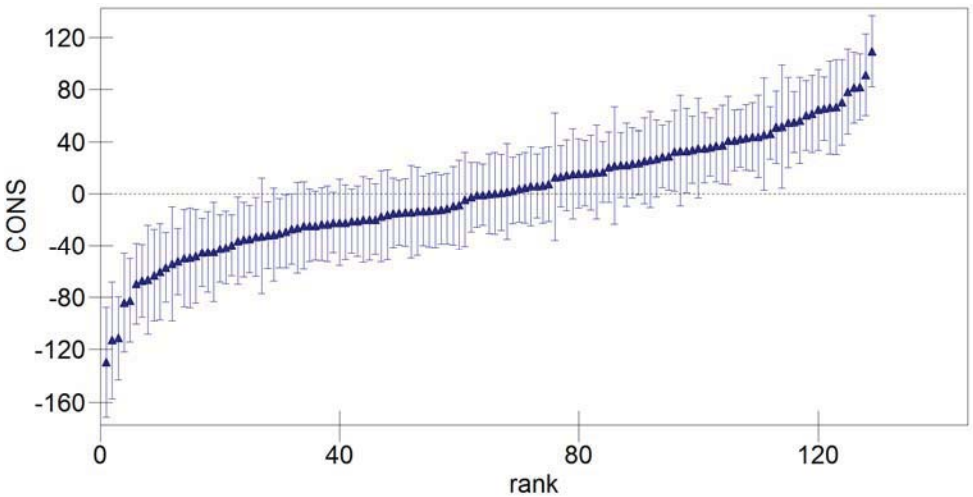
Estimation	Unconditional Model A_1 (IGLS)		Unconditional Model A_2 (MCMC 350K)	
Response: BSSSCIPV	Coefficient	s.e.	Coefficient (VPC)	s.e.
Fixed Part				
CONSTANT β_0	494.4	4.3	494.1	4.3
Random Part				
Level-4: IDSCHOOL f_{0l}	2118.9	296.9	2167.5 (0.33)	316.7
Level-3: IDTEACH v_{0kl}	492.1	74.6	491.1 (0.07)	73.9
Level-2: IDSTUD u_{0jkl}	3282.6	87.1	3286.7 (0.49)	87.2
Level-1: IDCASE e_{0ijkl}	706.7	8.1	706.8 (0.11)	8.1

Note: N =18845 cases, 3769 students, with 777 teachers, in 129 schools.

The model's diagnostics are evaluated in the same way as presented earlier, taking account of the accuracy levels as determined by the chain length and checking residual deviations for normality to confirm the residual variances are homogeneous. Much longer chains are required to overcome issues with auto-correlation and to satisfy accuracy levels for Raftery-Lewis quantiles and Brooks-Draper mean estimates. The point estimate can be cited to 3 significant figures on the basis of Brooks-Draper mean (Nhat = 56,331), but for anything more precise I need to run an MCMC with 800K iterations or more in some of the models. The kernel density estimate of the posterior distribution looks to be approximately following a Normal distribution as required for further analyses, making the point-estimates from IGLS analysis reasonably accurate but failing to fully provide an accurate breakdown of the variance components as referenced in section 5.3.

Inspection of the 'caterpillar' plots of residuals +/- 1.96 s.d. by rank for each level of the hierarchy shows the science data from G8 (secondary school education) is different from the equivalent data at G4 (primary school science education presented in 5.1). At school-level, there are 30-35 schools at each end of the plot shown to be significantly different from the predicted average. This visual representation reinforces the message that around one third of the total variance is attributed to school differences; these are the cases that can provide information on variant practices through analysis of the significantly different patterns of variance in the model. By comparison, there were very limited differences between G4 schools with only one case at either end of the distribution showing evidence of being significantly different from the predicted average i.e. not overlapping with zero.

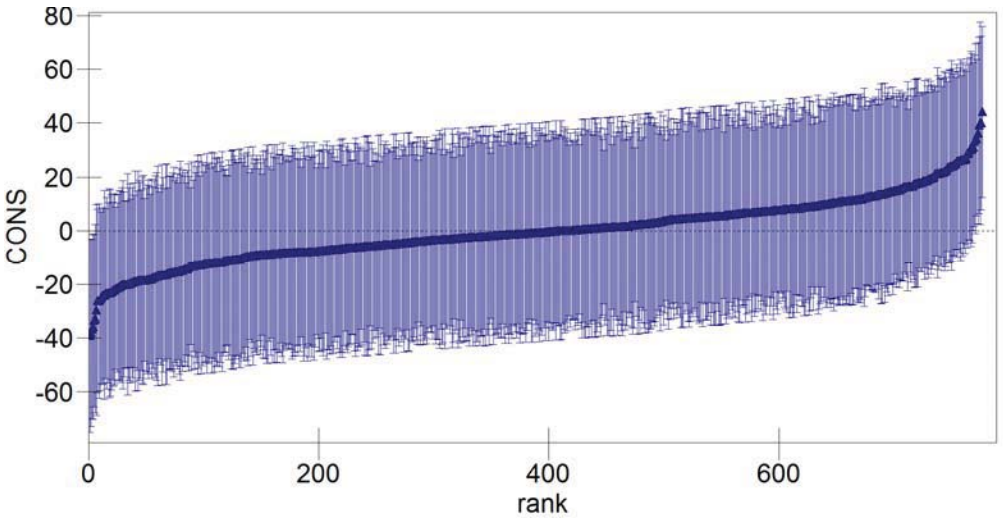
Figure 5.4-1: GR8 Science school-level residuals



The residuals at teacher-level of the models presented in Figure 5.4-2 shows there are very few cases where G8 teacher units are significantly different from the overall predicted score; only around five teachers at either end of the plot. The G4 data presented a

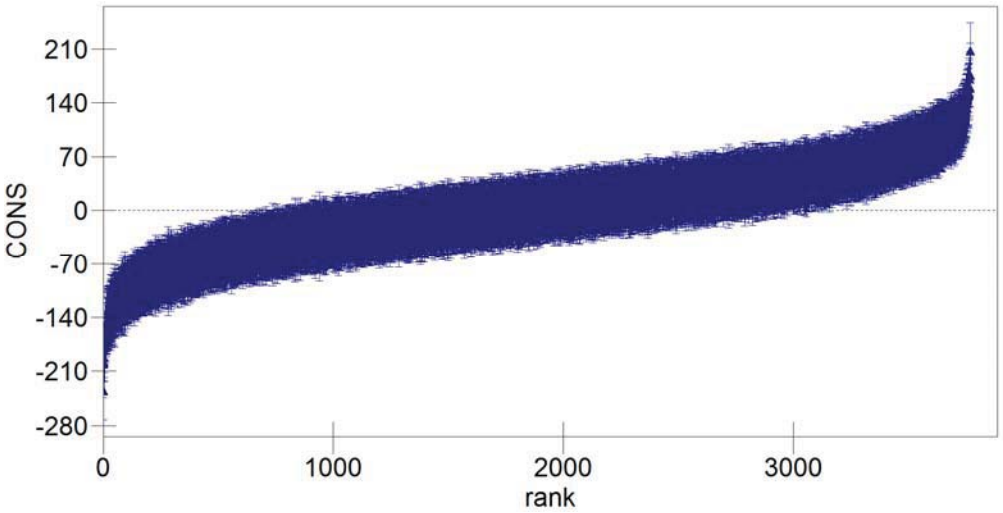
similar picture with only marginally more teacher cases providing residuals that are significantly different (at the 5% level) from the average predicted score for their school.

Figure 5.4-2: GR8 Science teacher-level residuals



The residuals at student-level of the GR 8 science data, as illustrated in Figure 5.4-3, are smaller than those reported at GR 4 and there are considerably fewer cases that can provide explanation of variance in the model: only around 630 cases in lower tail and 560 in upper tail, compared with the G4 plot that offers around 800 and 830 in the respective tails.

Figure 5.4-3: GR8 Science student-level residuals



There is scope to identify explanations for differences in achievement scores by looking to reduce the between-school and between-student variances through modelling control and predictor variables as pursued in earlier sections. The aim is to identify practices and experiences that are associated with achievement in order to inform policy and practice.

5.4.1. Control variables plus 'memorising science facts and principles'

In considering which control variables to use for the models of science achievement for G8, I started with the same set of variables identified in Chapter 5.2.1. First, the control variables for G8 data are not identical to those used earlier. Although the selection and structure of control variables for G8 is set out in Chapter 5.2, some of those variables do not provide significant explanation of variance in the multi-level models of G8 science achievement. As a result, some categorical variables were dropped from consideration; this removed possession of mobile phone (BS4GTH07) and the variable on safety concerns (BS4GPBSS) from subsequent models. The derived variables for out-of-school interests, categorised as 'Social', 'Individual' and 'Personal Development', are entered in each model using the pseudo-continuous scaled version of the variables described in Chapter 4.2.2. It is worth noting that the clusters under the same names are slight variants on those used with G4, with the 'social' dimension now based on two variables: play and talk with friends, and using the internet; the 'individual' component is also based on two responses: watch TV or videos, and play computer games; and the 'personal development' component is as before with three contributing elements: do jobs at home, read book for enjoyment, and do homework. These new clusters better reflect the G8 student responses. In the EDA I noted that there was a curvilinear association between those scaled variables and achievement scores; during this primary analysis the best fitting non-linear model was determined by fitting higher powers and reducing the complexity if no significant contribution to the model was made. The resultant framework includes quadratic models for the 'Social', 'Individual' and 'Personal Development' control variables.

The following set of control variables are therefore included as fixed factors in each of the models. The categorical variables of interest are presented with reference category as noted in parenthesis:

1. Speaks own language at home (always English)
2. ICT at home (no computer)
3. Books at home (one bookcase of 26-100 books)
4. Possessing own bedroom (Yes)
5. Work in paid job (no time)
6. Play sports (no time)
7. Gender (boy)
8. Formative years in UK (born or arrived in UK under 5 years)
9. Study tools (none of calculator, dictionary, encyclopaedia)
10. Out-of-school 'Social' (quadratic)
11. Out-of-school 'Individual' (quadratic)
12. Out-of-school 'Personal Development' (quadratic)

The review of science education in Chapter 2 guided the selection of predictor variables and focus on practices that explain differences within achievement scores. The explanatory factors used for G8 classroom experiences as are as set out in Table 5.1-1.

The first of those will be presented in full with summary data provided for the remaining predictors. The reported coefficients and standard errors associated with each of the control variable responses is fairly consistent across the models so the data in Table 5.4-2 is taken as representative of the control-effects and association with science achievement for G8 students.

Table 5.4-2: Memorise Science (Model B) includes control variables

MCMC Estimation (250,000 iterations) Response: BSSSCIPV	Unconditional Model A		Memorise Science Facts & Principles (Model B)		
	Coef.	s.e.	Coef.	s.e.	Sig.
Fixed Part					
CONSTANT β_0	495.2	4.3	474.9	9.3	**
MEMORISE SCIENCE FACTS & PRINCIPLES (NEVER)					
SOME LESSONS			9.1	3.6	*
ABOUT HALF THE LESSONS			16.1	3.6	**
EVERY OR ALMOST EVERY LESSON			12.5	3.8	**
FORMATIVE YEARS (BORN/UNDER					
ARRIVED IN UK AGED 5 TO 10 YEARS			-10.0	9.0	-
ARRIVED IN UK OLDER THAN 10			-30.3	6.8	**
SPEAKS ENGLISH AT HOME					
ALMOST ALWAYS			-4.8	3.9	-
SOMETIMES			-10.6	5.7	-
NEVER			-30.9	9.8	**
ICT AT HOME: COMPUTER, INTERNET(NONE)					
COMPUTER WITHOUT INTERNET			8.2	6.7	-
COMPUTER WITH INTERNET			22.4	5.9	**
BOOKS AT HOME (26 TO 100 BOOKS)					
NONE OR VERY FEW (0 TO 10			-35.7	2.9	**
ONE SHELF (11 TO 25 BOOKS)			-13.5	2.5	**
TWO BOOKCASES (101 TO 200			13.6	2.9	**
THREE + BOOKCASES (OVER 200 BOOKS)			34.8	2.9	**
POSSESS OWN BEDROOM (YES)					
NO			-7.0	2.6	**
WORK IN PAID JOB (NO TIME)					
LESS THAN 1 HOUR			-13.3	3.4	**
BETWEEN 1 TO 2 HOURS			-12.7	3.5	**
MORE THAN 2 BUT LESS THAN 4 HOURS			-11.3	4.4	*
4 OR MORE HOURS			-10.2	4.3	*
PLAY SPORTS (NO TIME)					
LESS THAN 1 HOUR			0.2	3.1	-
BETWEEN 1 TO 2 HOURS			3.9	3.0	-
MORE THAN 2 BUT LESS THAN 4 HOURS			0.9	3.3	-
4 OR MORE HOURS			-14.6	3.3	**

MCMC Estimation (175K iterations) Response: BSSSCIPV	Unconditional Model A Coef.	s.e.	Memorise Science Facts & Principles (Model B) Coef.	s.e.	Sig.
GENDER (BOY)					
GIRL			-14.1	2.1	**
STUDY TOOLS					
ONE OF CALC, DICT,			7.0	6.0	-
TWO OF CALC, DICT,			15.6	5.4	**
ALL THREE OF CALC, DICT,			19.4	5.4	**
OUT-OF-SCHOOL INTERESTS - SOCIAL					
(SCLSOCIAL-gm)^1			-15.1	2.5	**
(SCLSOCIAL-gm)^2			-7.8	2.4	**
OUT-OF-SCHOOL INTERESTS - INDIVIDUAL					
(SCLIND-gm)^1			4.7	1.8	**
(SCLIND-gm)^2			-8.4	1.7	**
OUT-OF-SCHOOL INTERESTS – PERSONAL					
(SCLPDEV-gm)^1			19.7	1.5	**
(SCLPDEV-gm)^2			-13.9	1.4	**

Random Part			Variance (VPC)		*Total Proportion Reduction*
Level-4: IDSCHOOL f_{0l}	2146.3	312.6	860.6 (0.20)	141.0	0.60
Level-3: IDTEACH v_{0kl}	486.9	74.9	283.7 (0.07)	53.9	0.42
Level-2: IDSTUD u_{0jkl}	3280.9	87.9	2441.2 (0.57)	67.6	0.26
Level-1: IDCASE e_{0ijkl}	705.7	8.2	705.7 (0.16)	8.2	0.00
Total Variance	6619.9		4291.3		0.35
DIC:	178269		178223.9	(45.2)	
pD:	3579.0		3533.5	(45.5)	

Note: N=18590 cases, 3679 students, with 772 teachers, in 129 schools;
** t-value>2.58, sig. level 1%; * t-value>1.96, sig. level 5%

All the control variables provide a significant explanation of variance in Model B. The overall reduction of 35% in total variance is partitioned with significant reductions in school- and teacher-level variance, reported as 60% and 42% respectively. A smaller proportion reduction of variance at student-level still leaves the bulk of unexplained variance (57%) to be at student-level with one fifth of unexplained variance at school-level and only a small proportion (7%) remaining at teacher-level of the model. Model B provides a good fit to the data with reductions in DIC and in model complexity with a simpler structure reported as we move from the unconditional model.

The bulk of the control variables carry negative coefficients, highlighting the

limiting or qualifying effect relative to the reference categories that those experiences can have on students' achievement in scientific studies. The variables associated with the greatest effect on achievement are related to home background: the number of books at home; whether English is ever spoken at home; migrant family having only recently arrived in UK (aged 10 or over on arrival); home support through learning study tools (calculator, dictionary, and encyclopaedia); internet connection for home computer. Possession of high quality support is clearly associated with a positive effect, whereas a shortfall carries a negative association with achievement. Remembering these students are in G8 and only aged 13/14, it is not unsurprising that those who undertake any paid work, even for only a short period of time in the week, are associated with a negative effect on academic achievement in science.

5.4.2. Other predictor variables (G8 science)

The models of science achievement scores for the other predictor variables are presented in Table 5.4-3. Each model is presented in summary format, acknowledging the same systematic approach as used with the development of previous models was deployed *en route* to the data presented here. The table presents the coefficients, standard errors, and measures of significance for the predictor variable from the univariate model and its associated multivariate cluster. The multivariate analysis is based on the cluster of variables within the selected theme, noting resultant statistics in parenthesis where they differ from the univariate model. The diagnostics of each univariate model and partitioning of variance within those models is presented in Appendix 5.3.1. The key data in Table 5.4-3 will guide discussion of findings, prioritising as it will the experiences that offer greatest explanation of variance in science achievement, supporting conclusions drawn on which aspects of learning and teaching, and classroom experience, appear to offer explanation of variance in the models of student's science achievement.

Table 5.4-3: Summary of Effect Sizes (Models of G8 Science Achievement)

MCMC Estimation	*Univariate Model Coefficients & Significance (Multivariate)*			*Univariate Effect size (Multivariate)*	
Response: BSSSCIPV	Coef.	s.e.	Sig.	Δ	+/- error
MODEL B: EXPLAIN SCIENCE STUDIED (NEVER)					
SOME LESSONS	4.9 (3.3)	4.1 (4.2)	-	0.09 (0.06)	0.07
ABOUT HALF THE LESSONS	13.2 (10.4)	4.1 (4.3)	** (*)	0.23 (0.18)	0.07 (0.08)
EVERY OR ALMOST EVERY LESSON	15.8 (13.1)	4.1 (4.4)	**	0.28 (0.23)	0.07
MODEL C: MEMORISE SCIENCE FACTS (NEVER)					
SOME LESSONS	7.3 (5.0)	3.8 (3.9)	-	0.13 (0.09)	0.07
ABOUT HALF THE LESSONS	13.9 (9.1)	3.8 (4.0)	** (*)	0.25 (0.16)	0.07 (0.08)
EVERY OR ALMOST EVERY LESSON	12.2 (5.9)	4.0 (4.3)	** (-)	0.22 (0.10)	0.07 (0.08)
MODEL D:WATCH TEACHER DEMONSTRATE (NEVER)					
SOME LESSONS	6.2 (0.9)	7.0 (7.4)	-	0.11 (0.02)	0.12 (0.13)
ABOUT HALF THE LESSONS	15.7 (7.8)	6.9 (7.4)	* (-)	0.28 (0.14)	0.12 (0.13)
EVERY OR ALMOST EVERY LESSON	8.6 (1.4)	6.9 (7.4)	-	0.15 (0.02)	0.12 (0.13)
MODEL F: CONDUCT AN EXPERIMENT (NEVER)					
SOME LESSONS	14.2 (12.8)	6.9 (7.4)	* (-)	0.25 (0.23)	0.12 (0.13)
ABOUT HALF THE LESSONS	24.2 (20.9)	6.9 (7.4)	**	0.43 (0.37)	0.12 (0.13)
EVERY OR ALMOST EVERY LESSON	20.8 (21.0)	6.9 (7.5)	**	0.37	0.12 (0.13)

MCMC Estimation	*Univariate Model Coefficients & Significance (Multivariate)*			*Univariate Effect size (Multivariate)*	
Response: BSSSCIPV	Coef.	s.e.	Sig.	Δ	+/- error
MODEL E: DESIGN OR PLAN AN EXPERIMENT (NEVER)					
SOME LESSONS	5.6 (2.6)	4.1 (4.3)	-	0.10 (0.05)	0.07 (0.08)
ABOUT HALF THE LESSONS	10.1 (3.2)	4.1 (4.4)	* (-)	0.18 (0.06)	0.07 (0.08)
EVERY OR ALMOST EVERY LESSON	4.5 (-1.7)	4.2 (4.7)	-	0.08 (-0.03)	0.07 (0.08)
MODEL G: WORK ON OWN(NEVER)					
SOME LESSONS	4.3 (3.3)	3.7	-	0.08 (0.06)	0.06 (0.07)
ABOUT HALF THE LESSONS	10.8 (6.9)	3.7	** (-)	0.19 (0.12)	0.07
EVERY OR ALMOST EVERY LESSON	10.5 (5.9)	3.9 (4.0)	** (-)	0.19 (0.10)	0.07
MODEL H: WORK IN SMALL GROUP (NEVER)					
SOME LESSONS	11.3 (7.0)	8.1 (8.6)	-	0.20 (0.12)	0.14 (0.15)
ABOUT HALF THE LESSONS	22.9 (15.4)	8.0 (8.6)	** (-)	0.41 (0.27)	0.14 (0.15)
EVERY OR ALMOST EVERY LESSON	20.2 (11.3)	8.0 (8.5)	* (-)	0.36 (0.20)	0.14 (0.15)
MODEL L: LECTURE-STYLE PRESENTATION					
SOME LESSONS	6.4 (3.7)	6.4 (6.8)	-	0.11 (0.06)	0.11 (0.12)
ABOUT HALF THE LESSONS	15.4 (9.8)	6.3 (6.7)	* (-)	0.27 (0.17)	0.11 (0.12)
EVERY OR ALMOST EVERY LESSON	23.4 (18.2)	6.1 (6.5)	**	0.41 (0.32)	0.11 (0.12)
MODEL J: REVIEW HOMEWORK (NEVER)					
SOME LESSONS	-3.0 (-3.7)	2.5	-	-0.05 (-0.07)	0.04
ABOUT HALF THE LESSONS	-7.4 (-6.2)	2.8	** (*)	-0.13 (-0.11)	0.05
EVERY OR ALMOST EVERY LESSON	-11.6 (-7.4)	2.9 (3.0)	** (*)	-0.21 (-0.13)	0.05
MODEL K: QUIZ OR TEST (NEVER)					
SOME LESSONS	11.5 (12.5)	6.0	- (*)	0.20 (0.22)	0.11
ABOUT HALF THE LESSONS	2.0 (4.1)	6.3 (6.4)	-	0.03 (0.13)	0.11
EVERY OR ALMOST EVERY LESSON	-14.2 (-11.6)	6.6 (6.7)	* (-)	-0.25 (-0.21)	0.12
MODEL M: COMPUTER (NEVER)					
SOME LESSONS	3.8 (2.1)	2.1 (2.2)	-	0.07 (0.04)	0.04
ABOUT HALF THE LESSONS	-12.9 (-12.3)	3.7 (3.8)	**	-0.23 (-0.22)	0.07
EVERY OR ALMOST EVERY LESSON	-30.2 (-28.1)	5.5 (5.6)	**	-0.53 (-0.50)	0.10
MODEL N: COMPUTER FOR SCH WORK (NO TIME)					
A FEW TIMES A YEAR	11.2 (10.9)	2.4 (2.5)	**	0.20 (0.19)	0.04
ONCE OR TWICE A MONTH	5.1 (5.6)	2.5	* (*)	0.09 (0.10)	0.04
AT LEAST ONCE A WEEK	-10.2 (-8.3)	3.5 (3.6)	** (*)	-0.18 (-0.15)	0.06
EVERY DAY	-21.7 (-16.6)	7.1	** (*)	-0.38 (-0.29)	0.13
MODEL P: DAILY LIFE (NEVER)					
SOME LESSONS	-2.3	2.9	-	-0.04	0.05
ABOUT HALF THE LESSONS	3.0	3.0	-	0.05	0.05
EVERY OR ALMOST EVERY LESSON	1.7	3.3	-	0.03	0.06

** t-value>2.58, sig. level 1%; * t-value>1.96, sig. level 5%

Active learning and practical activities

The predictor variable with highest effect in these models is the one that relates to 'conducting' science experiments or investigations (Model F, equivalent to 'doing' at G4), where a strong association with achievement scores in science is highlighted. Two of the three response categories offer significant explanation of variance in the multivariate model with effect size of $\Delta = 0.37$, equating to students exceeding the achievement scores of 64% of the control group.

The combination of 'conducting' experiments with opportunities to 'watch' science experiments or investigations provided a significant improvement in the model as a whole, as illustrated in Model O (Appendix 5: Table 6.5-14), where a loglikelihood reduction of 16.9 is highly significant with $p<0.01$ for CPRObability (16.9, 3). This finding reinforced the dominant benefit of conducting experiments, but watching teachers demonstrate experiments lost significance in the multivariate model. Incorporating 'planning' into the model provided no further explanation of variance in Model P (Table 6.3-14 in Appendix 5). In summary, higher science achievement scores are associated with regular opportunities for students to engage with practical scientific enquiry through conducting experiments or investigations; no additional explanation or association with achievement is attributed to watching teacher demonstrations or having opportunities to plan science experiments or investigations.

Learning environment

Although the univariate models provided significant evidence of effect in their respective models of science achievement, when the variables were combined the only significant contributor in this multivariate cluster concerns opportunities for lecture-style presentation. There is a high effect of $\Delta = 0.32$ reported when students experience lecture-style in every or almost every lesson. The magnitude of effect is equivalent to those students exceeding the achievement scores of over 63% of the control group. Other aspects of the learning environment provide no significant evidence of explanation in the multivariate model of science achievement, although there are higher effects attributed to working in a small group (maximum $\Delta = 0.27$) than working on own (maximum $\Delta = 0.12$).

Reform-based practice and discussion

The model coefficients in Table 5.4-3 show changes in the levels of significance for the multivariate model when compared with the univariate models, with a strengthening of effect attributed to explaining science studied. Two of the response categories provide

significant explanation of variance whereas only one response category of the competing concept of 'memorising' retains significance. The strongest effect is associated with explaining answers on 'every or almost every lesson' ($\Delta = 0.23$), equivalent to those students exceeding the achievement scores of 59% of the control group.

Technology

The role of ICT in learning and teaching G8 science does not offer a positive association with achievement. The response to using a computer in science lessons (every or almost every lesson) is negatively associated with achievement and is attributed the highest absolute effect size across all models of science achievement considered in this stage ($\Delta = -0.50$); same direction of association but smaller effect ($\Delta = -0.22$) is noted for the response category of using computers in about half of the science lessons. There is evidence of small but significant benefits in *occasionally* using computers for school work (in and out of school), with an effect size of $\Delta = 0.19$ when experienced 'a few times a year'. However, as frequency of computer use increases, the strength of association with achievement diminishes before showing a significant negative association when students report this type of activity 'every day' ($\Delta = -0.29$).

Assessment and feedback

One variable in this cluster is positively associated with science achievement scores, namely having a quiz or test *occasionally* i.e. in some lessons, where $\Delta = 0.22$. Other response categories on quiz or test are negatively associated with achievement but are non-significant within the multivariate model that jointly analyses Model J: Review Homework and Model K: Quiz or Test. All response categories of reviewing homework are associated with small negative effects that are significant contributors when the frequency extends beyond half the lessons ($\Delta = -0.11\ to\ -0.13$).

Contexts (real life)

Much as noted in the G4 analysis, modelling the variable on relating learning to daily life provides no explanation of variance in the G8 science achievement data. There is a very equitable split across the response categories with minimal variance in mean achievement scores reported in EDA. None of the parameter estimates calculated through MCMC method are significant. This variable was consequently dropped from clusters but included in final analysis that contained all of the predictor variables as guided by the literature.

5.4.3. Modelling all predictor and control variables (G8 science)

Combining all background control variables and predictor variables will allow competing themes within the data and to be evaluated. This model, an extension of groups of related variables reported in Table 5.4-3 will help identify the principal associations with G8 science achievement.

The findings are presented in full in Appendix 5.3.2. In summary, the findings from this model are presented in Table 5.4-4 where positive and negative association with science achievement are presented against predictor responses.

Table 5.4-4: Summary of key findings in model of G8 science achievement data with all control and predictor variables included

Response	Effect size	
	Negative association	Positive association
Explain science studied(never)		
Every or almost every lesson		0.22
Lecture-style presentation (never)		
Every or almost every lesson		0.25
Work problems on our own (never)		
Every or almost every lesson		0.19
Review our homework (never)		
About half the lessons	-0.23	
Every or almost every lesson	-0.28	
Have a quiz or test (never)		
Every or almost every lesson	-0.25	
Use computers in sci lesson(never)		
About half the lessons	-0.18	
Every or almost every lesson	-0.36	
Computer in support of sch wk(none)		
A few times a year		0.16
At least once a week	-0.17	
Every day	-0.28	

In secondary science education it is apparent that opportunities to explain science studied are strongly associated with achievement. This presents a different picture to that reported at G4, where memorisation was a dominant contributor in the model of science achievement; memorisation at G8 loses significance in the combined model, much as predicted within the multivariate analyses in Table 5.4-3: Summary of Effect Sizes (Models of G8 Science Achievement). Around one third of the cohort report explaining science on every or almost every lesson, the response category that is significantly associated with high achievement ($\Delta = 0.22$).

The key features of students' learning environment that retained a positive association with science achievement in the combined model are lecture-style presentation and working on own. Regular lecture-style presentations provide the stronger effect, with $\Delta = 0.25$ being equivalent to those students out performing sixty per cent of the control group. Working on own is a more significant contributor (t-value > 2.58) but the response category of every or almost every lesson has a smaller effect in the combined model with $\Delta = 0.19$; working in small groups offers a similar effect size ($\Delta = 0.19$) but this is a non-significant contributor in the combined model.

Although there is a small positive association reported for computer use in support of schoolwork 'a few times a year' ($\Delta = 0.16$), the general pattern for the role of ICT in support of science school work, in or out of lessons, is an increasingly negative association with achievement. As the frequency of use in class rises to every or almost every lesson, the effect size increases to $\Delta = -0.36$; equivalent to 64% of the control group outperforming those students. Similarly for out-of-school computer use in support of science school work , a significant negative association with achievement is reported when this occurs at least once a week $\Delta = -0.17$, increasing to $\Delta = -0.28$ for computer use 'every day'.

In the cluster of variables within assessment and feedback, the response categories that provide a significant explanation in the model are all negatively associated with science achievement. The highly significant effects associated with reviewing homework in about half the lessons rising to every or almost every lesson, go from $\Delta = -0.23$ to $\Delta = -0.28$. Those response categories and effects impact almost fifty per cent of the cohort. A comparable negative association ($\Delta = -0.25$) is noted for having a quiz or test every or almost every lesson. Admittedly this only affects fewer than ten per cent of the cohort but it is striking that feedback accrued from quiz, test or reviewing homework does not provide a significant positive association with science achievement.

5.5. Mathematics Education (G8)

The general multilevel model for G8 mathematics achievement has an identical structure to that described in 5.3, with 'plausible value' taken as the lowest level (*i*) in the model, nested within student (*j*), teacher (*k*), and school (*l*) levels of analysis. The form of the general model of mathematics achievement is written as:

$$BSMMATPV_{ijkl} = \beta_0 + f_{0l} + v_{0kl} + u_{0jkl} + e_{0ijkl}$$

where G8 mathematics achievement is the response variable and β_0 is the overall mean score across all schools. The error in student achievement is presented separately for each level of the model as before, providing a breakdown of error in the model as a whole. These are shown in the unconditional models for the G8 students in Table 5.5-1.

Table 5.5-1: Unconditional Model of GR8 Mathematics Achievement

Estimation		Unconditional Model A_1 (IGLS)		Unconditional Model A_2 (MCMC 60,000 iterations)		Unconditional Model A_3 (MCMC 350,000 iterations)	
Response: BSMMATPV		Coefficient	s.e.	Coefficient	s.e.	Coefficient	s.e.
Fixed Part							
CONSTANT	β_0	476.2	4.3	476.0	3.8	476.0	4.0
Random Part							
Level-4: IDSCHOOL	f_{0l}	573.5	309.5	92.2	262.4	349.2	360.0
Level-3: IDTEACH	v_{0kl}	3792.4	399.5	4304.1	423.8	4059.0	463.1
Level-2: IDSTUD	u_{0jkl}	1451.6	36.0	1451.7	36.3	1452.1	36.2
Level-1: IDCASE	e_{0ijkl}	507.8	5.7	507.8	5.6	507.9	5.7
Total Variance		6325.3		6355.7		6368.1	
-2*LOGLIKELIHOOD		195011.8					
DIC:				186741.3		186739.6	
pD:				3792.4		3791.4	

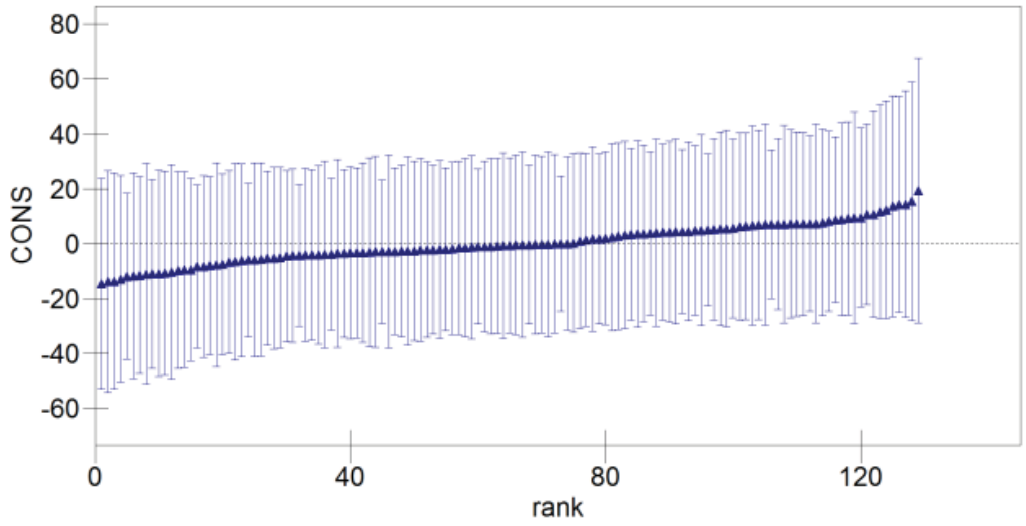
Note: N =20175 cases, 4035 students, with 314 teachers, in 129 schools;

These estimates are computed using different techniques, with Model A_1 generated via Iterative Generalised Least Squares method (IGLS) and Model A_2 derived from a Monte Carlo Markov Chain method (MCMC) as used throughout previous sections. The observed differences are in the analysis of variance in the hierarchical structure of the model. The MCMC methodology provides a more accurate breakdown of the variance components as the number of iterations increases, with diagnostic checks providing support for the accuracy of estimates. As in section 5.3, the diagnostic checks are used to confirm assumptions on variance components, to check residual deviations are normally distributed and to confirm

residual variances are homogeneous. There are no threats to validity in the model.

The partitioning of variance within the model is quite different to that reported on G4 data. The bulk of variance in the G8 data is at teacher-level (64%) with a small proportion attributed to school-level (5%) and over a fifth at student-level of the hierarchy (23%); some 8% variance is contained within the plausible values at unit-level. This compares with VPCs of 0.12, 0.07, 0.69 and 0.12 respectively within the G4 data in section 5.3. This shift in proportions of variance will be primarily down to organisational structures within the secondary school settings, where the classes are likely to be set by ability in contrast to the primary school arrangement of mixed ability groupings; however this is not explicitly recorded in the data and is therefore a speculative interpretation of the findings. Assuming the classes are set by ability then differences between classes in each school (teacher variance) will be large, and differences within classes (student variance) will be small. Inspection of the 'caterpillar' plots of residuals +/- 1.96 s.d. by rank for each level of the hierarchy provides a visual of the data from G8, reinforcing the differences from the G4 data presented in section 5.3 and perhaps reflecting the organisational differences between sectors. At school-level, the G8 spread of scores within each school is such that no schools are identified as having residuals that are significantly different from the overall mean; there are no significant differences between schools. In G4 the pattern of residuals was more pronounced but there was only one school where the plot did not overlap with zero, marking the residuals as significantly different (at the 5% level) from β_0 the overall average predicted score.

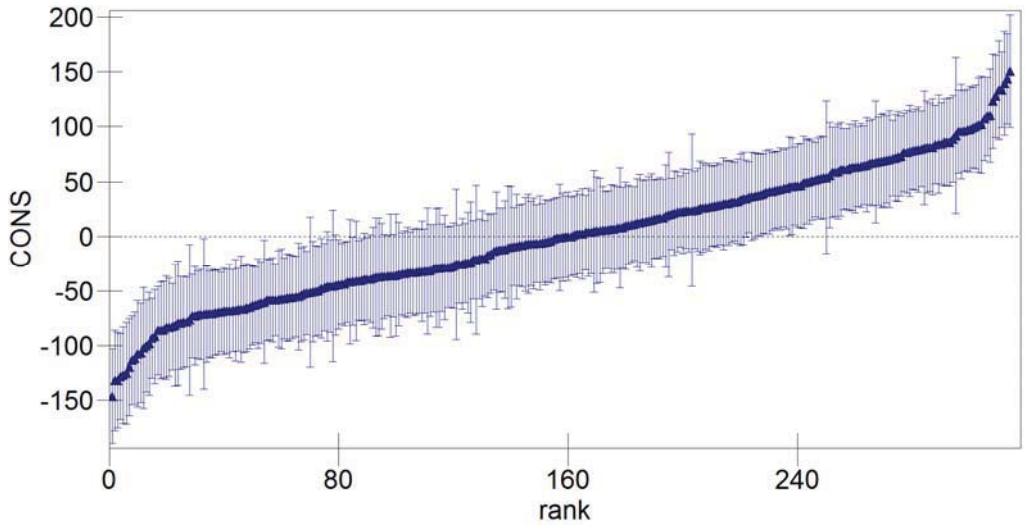
Figure 5.5-1: Caterpillar Plots of GR8 school-level residuals



The teacher-level of the model provides the greatest explanation for differences, with over ninety cases in each tail of the plot reflecting the number of teachers with residuals

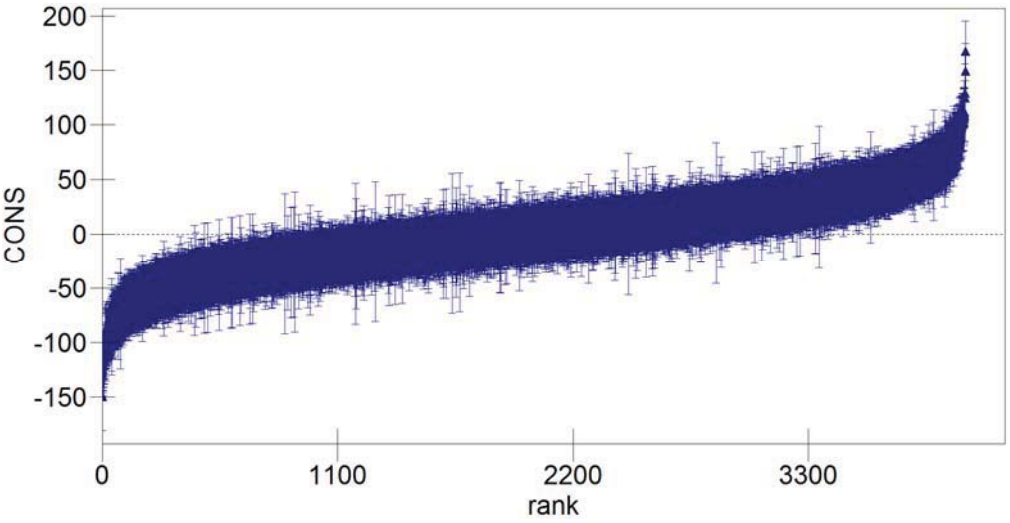
that were significantly different from the predicted school average. This represents a major shift from that witnessed in the G4 data where fewer than ten teachers in each tail of the caterpillar plot displayed significance.

Figure 5.5-2: Caterpillar Plots of GR8 teacher-level residuals



The residuals at student-level of the G8 data are generally smaller than those reported for G4 and do not provide the same degree of significant explanation of variance within the model. This is evidenced via the VPC computed from Table 5.5-1 but also through the student-level residual plot by rank shown below, where there are around 700 cases in each of the lower and upper tails of the plot that do not overlap with zero. This compares with around 870 cases in the tails of G4 Mathematics student-level plot.

Figure 5.5-3: Caterpillar Plots of GR8 student-level residuals



There is therefore ample scope to identify plausible explanations for differences in

achievement scores by looking to reduce the between-teacher and between-student variances by modelling control and explanatory variables as pursued in previous sections. The exploratory data analysis detailed in Chapter 4 and the review of mathematics education in Chapter 2, guided the selection of control variables and pedagogical focus on experiences that explain differences within achievement scores.

The control variables for G8 Mathematics data are not identical to those used earlier. The categorical variables used in the G8 Science models were all considered for inclusion but some were dropped to make a more parsimonious model because they did not provide any significant explanation in the model. This led to the removal of variables reporting on study tools, ICT at home, possession of own bedroom, possession of mobile and the variable on safety concerns, none of which provided explanation of variance in the models. The derived variables for out-of-school interests categorised as 'Social', 'Individual', and 'Personal Development' dimensions are entered in each model using the pseudo-continuous scaled version of the variables outlined in Chapter 4.2.2. It is worth noting that the clusters are derived on the same basis as for Science Education, with the empirical analysis used to determine the best fitting non-linear model by fitting higher powers and reducing the complexity if no significant contribution was made. The resultant framework includes quadratic models for the 'Social' and 'Individual' variables, and a cubic model for the 'Personal Development' control variable. The control variables to be included in models of mathematics achievement scores are:

1. Speaks own language at home (always English)
2. Books at home (one bookcase 26-100 books)
3. Formative years in UK (born or arrived in UK under 5 years)
4. Gender (boy)
5. Work in paid job (no time)
6. Play sports (no time)
7. Out-of-school 'Social' (quadratic)
8. Out-of-school 'Individual' (quadratic)
9. Out-of-school 'Personal Development' (cubic)

5.5.1. Control variables plus 'memorising formulas and procedures'

The selected control variables all provide significant explanation of variance in the models of student achievement. Typical values for coefficients, standard errors and effect sizes are illustrated in Table 5.5-2, where the tabulated values for control variables are computed from data based on the 'memorise' model. These reported values are from the MCMC estimation of student achievement but it is worth noting that the IGLS estimation produced virtually identical data; IGLS methodology will be used in development of later models. The data from Model B in Table 5.5-2 are used to summarise the effects of control variables in the models.

Table 5.5-2: Memorising procedures and formulae (Model B) compared with Unconditional (Model A′)

MCMC Estimation (54,000 iterations)	*Unconditional Model (Model A')*		*Memorise Procedures & Formulae (Model B)*		
Response: BSMMATPV	Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part					
CONSTANT β_0	477.1	3.8	487.1	4.3	**
MEMORISE PROCEDURES ETC. (NEVER)					
SOME LESSONS			9.8	2.1	**
ABOUT HALF THE LESSONS			9.9	2.2	**
EVERY OR ALMOST EVERY LESSON			10.2	2.4	**
SPEAKS ENGLISH AT HOME					
ALMOST ALWAYS			-0.2	2.9	-
SOMETIMES			5.8	4.1	-
NEVER			-13.9	7.0	*
BOOKS AT HOME (26 TO 100 BOOKS)					
NONE OR VERY FEW (0 TO 10			-16.2	2.0	**
ONE SHELF (11 TO 25 BOOKS)			-9.0	1.8	**
TWO BOOKCASES (101 TO 200 BOOKS)			2.9	2.1	-
THREE + BOOKCASES (OVER 200 BOOKS)			12.7	2.1	**
FORMATIVE YEARS (BORN/UNDER 5)					
ARRIVED IN UK AGED 5 TO 10			3.8	7.0	-
ARRIVED IN UK OLDER THAN 10			-17.8	5.1	**
GENDER (BOY)					
GIRL			-10.0	1.5	**
WORK IN PAID JOB (NO TIME)					
LESS THAN 1 HOUR			-5.3	2.5	*
BETWEEN 1 TO 2 HOURS			-4.4	2.6	-
MORE THAN 2 BUT LESS THAN 4 HOURS			-8.4	3.2	**
4 OR MORE HOURS			-15.6	3.2	**

MCMC Estimation (54K iterations)	*Unconditional Model (Model A')*		*Memorise Procedures & Formulae (Model B)*		
Response: BSMMATPV	Coefficient	s.e.	Coef.	s.e.	Sig.
PLAY SPORTS (NO TIME)					
LESS THAN 1 HOUR			0.6	2.3	-
BETWEEN 1 TO 2 HOURS			3.4	2.2	-
MORE THAN 2 BUT LESS THAN 4 HOURS			-1.6	2.4	-
4 OR MORE HOURS			-7.1	2.4	**
OUT-OF-SCHOOL INTERESTS – SOCIAL					
(BSCLSOC-gm)^1			-7.3	1.3	**
(BSCLSOC-gm)^2			-4.2	1.3	**
OUT-OF-SCHOOL INTERESTS - INDIVIDUAL					
(BSCLIND-gm)^1			0.6	1.1	-
(BSCLIND-gm)^2			-3.2	1.1	**
OUT-OF-SCHOOL INTERESTS – PERSONAL					
(BSCLPD-gm)^1			2.4	2.0	-
(BSCLPD-gm)^2			-13.3	3.1	**
(BSCLPD-gm)^3			3.7	1.8	*
Random Part			*Variance (VPC)*		*Total Proportion Reduction*
Level-4: IDSCHOOL f_{0l}	229.2	276.2	158.9 (0.03)	208.1	0.31
Level-3: IDTEACH v_{0kl}	4092.7	426.4	3288.5 (0.63)	331.7	0.20
Level-2: IDSTUD u_{0jkl}	1429.0	37.0	1269.5 (0.24)	33.3	0.11
Level-1: IDCASE e_{0ijkl}	502.3	5.8	502.2 (0.10)	5.8	0.00
Total Variance	6253.2		5219.1		0.17
DIC:	172276.9		172254	(22.5)	
pD:	3501.9		3478.6	(23.3)	

Note: N=18635 cases, 3727 students, with 312 teachers, in 129 schools;
** t-value>2.58, sig. level 1%; * t-value>1.96, sig. level 5%

Model B has accounted for 17% of the unexplained variance in the unconditional model. The percentage reduction in variance computed at each level of Model B shows the inclusion of all the control variables and the predictor 'memorise' reduces the school-level variance by 31%. Teacher-level variance is reduced by 20% and the student-level variance is reduced by just over a tenth. Model B provides a good fit to the empirical data with reductions in DIC and model complexity (pD). The overall variance is now partitioned in such a way that the bulk of variance (63%) is still at teacher-level, with 24% remaining at student-level and only 3% of the unexplained variance at school-level.

5.5.2. Other predictor variables (G8 mathematics)

The G8 classroom experiences can be empirically evaluated in terms of learning and teaching mathematics and are considered as explanatory factors to generate separate models of mathematics achievement. The variables identified in section 5.1 are clustered by themes in Table 5.5-3, where each model is presented in summary format, acknowledging the same systematic approach as used within the development of previous models was deployed. Only the coefficients, standard errors, measures of significance, and effect sizes for the predictor variables B through N are presented; fuller details of each model are provided in Appendix 5.4.1. The same set of control variables as used in G8 Model B are added as fixed factors in each of the models.

Table 5.5-3: Summary of Effect Sizes (Models of Mathematics Achievement)

MCMC Estimation	*Univariate Model Coefficients & Significance (Multivariate where changes occur)*			*Univariate Effect size (Multivariate where changes occur)*	
Response: BSMMATPV	Coef.	s.e.	Sig.	Δ	+/- error
MODEL B: MEMORISE PROCEDURES ETC. (NEVER)					
SOME LESSONS	10.0 (9.5)	2.1 (2.2)	**	0.26 (0.25)	0.06
ABOUT HALF THE LESSONS	10.7 (10.0)	2.3 (2.4)	**	0.28 (0.26)	0.06
EVERY OR ALMOST EVERY LESSON	10.7 (10.0)	2.5	**	0.28 (0.27)	0.07
MODEL C: EXPLAIN MY ANSWERS (NEVER)					
SOME LESSONS	6.2 (4.4)	3.5	-	0.17 (0.12)	0.09
ABOUT HALF THE LESSONS	8.5 (6.2)	3.5	* (-)	0.23 (0.17)	0.09
EVERY OR ALMOST EVERY LESSON	7.9 (5.4)	3.3 (3.4)	* (-)	0.21 (0.14)	0.09
MODEL G:DAILY LIFE (NEVER)					
SOME LESSONS	-0.4	1.9	-	-0.01	0.05
ABOUT HALF THE LESSONS	0.5	2.1	-	0.01	0.06
EVERY OR ALMOST EVERY LESSON	-5.0	2.3	*	-0.13	0.06
MODEL H:REVIEW HOMEWORK (NEVER)					
SOME LESSONS	5.3 (5.2)	2.3	*	0.14	0.06
ABOUT HALF THE LESSONS	4.4 (5.4)	2.4	- (*)	0.12 (0.14)	0.06
EVERY OR ALMOST EVERY LESSON	-0.7 (1.1)	2.4	-	-0.02 (0.03)	0.06
MODEL I:QUIZ OR TEST (NEVER)					
SOME LESSONS	-10.6 (-11.2)	5.2	*	-0.28 (-0.31)	0.14
ABOUT HALF THE LESSONS	-17.7 (-18.4)	5.4	**	-0.47 (-0.49)	0.14
EVERY OR ALMOST EVERY LESSON	-24.1 (-23.6)	5.6	**	-0.64 (-0.63)	0.15

MCMC Estimation	*Model Coefficients & Significance*			*Effect size*	
Response: BSMMATPV	Coef.	s.e.	Sig.	Δ	+/- error
MODEL E: WORK IN SMALL GROUPS (NEVER)					
SOME LESSONS	-0.3 (-0.1)	1.6	-	-0.01	0.04
ABOUT HALF THE LESSONS	-1.9 (-3.3)	2.3	-	-0.05 (-0.09)	0.06
EVERY OR ALMOST EVERY LESSON	-15.7 (-18.7)	2.9	**	-0.42 (-0.50)	0.08
MODEL F: WORK ON OWN (NEVER)					
SOME LESSONS	4.8 (3.5)	3.5	-	0.13 (0.09)	0.09
ABOUT HALF THE LESSONS	7.1 (6.0)	3.4 (3.5)	* (-)	0.19 (0.16)	0.09
EVERY OR ALMOST EVERY LESSON	13.7 (13.6)	3.5	**	0.36 (0.36)	0.09
MODEL J:LECTURE-STYLE PRESENTATION (NEVER)					
SOME LESSONS	7.4 (7.6)	4.6 (4.5)	-	0.20	0.12
ABOUT HALF THE LESSONS	12.7	4.5	**	0.34	0.12
EVERY OR ALMOST EVERY LESSON	11.7 (10.5)	4.4	** (*)	0.31 (0.28)	0.12
MODEL K: CALCULATOR (NEVER)					
SOME LESSONS	8.7 (8.4)	4.4 (4.4)	* (-)	0.23 (0.2)	0.12
ABOUT HALF THE LESSONS	12.3 (13.0)	4.6	**	0.33 (0.34)	0.12
EVERY OR ALMOST EVERY LESSON	9.2 (11.5)	4.7	- (*)	0.24 (0.31)	0.13
MODEL L:COMPUTER (NEVER)					
SOME LESSONS	-2.5 (-3.0)	1.9	-	-0.07 (-0.08)	0.05
ABOUT HALF THE LESSONS	-12.6 (-13.0)	3.6 (3.7)	**	-0.33 (-0.34)	0.10
EVERY OR ALMOST EVERY LESSON	-15.2 (-14.8)	4.6	**	-0.40 (-0.39)	0.12
MODEL M: COMPUTER FOR SCH WK (NO TIME)					
A FEW TIMES A YEAR	4.8 (5.3)	1.8	**	0.13 (0.14)	0.05
ONCE OR TWICE A MONTH	-0.7 (-0.1)	1.9 (2.0)	-	-0.02 (0.00)	0.05
AT LEAST ONCE A WEEK	-3.8 (-2.6)	2.7	-	-0.10 (-0.07)	0.07
EVERY DAY	-28.4 (-27.1)	5.3 (5.4)	**	-0.75 (-0.72)	0.14
MODEL D: INTERPRET DATA (NEVER)					
SOME LESSONS	3.2 (2.1)	3.5	-	0.08 (0.06)	0.09
ABOUT HALF THE LESSONS	-4.0 (-5.4)	3.7 (3.8)	-	-0.10 (-0.14)	0.10
EVERY OR ALMOST EVERY LESSON	-11.8 (-13.6)	4.1 (4.2)	**	-0.31 (-0.36)	0.11
MODEL N: PROCEDURES FOR COMPLEX PROB (NEVER)					
SOME LESSONS	4.6 (5.1)	1.9	*(**)	0.12 (0.14)	0.05
ABOUT HALF THE LESSONS	2.0 (4.5)	2.2	- (*)	0.05 (0.12)	0.06
EVERY OR ALMOST EVERY LESSON	0.8 (5.3)	2.5 (2.6)	- (*)	0.02 (0.14)	0.07

** t-value>2.58, sig. level 1%; * t-value>1.96, sig. level 5%

The resultant models of mathematics achievement scores will be used to draw conclusions on which aspects of learning and teaching and classroom experience appear to offer explanation of variance within the hierarchical models of G8 mathematics achievement. In each model there is evidence of a good fit with some 15 to18% reduction in overall variance from the unconditional model. The control variables account for a significant proportion of that reduction in variance but the predictor variables offer additional explanation with significance and effect sizes as detailed in Table 5.5-3. The variance reduction is primarily attributed to school- and teacher-levels of the model with just under one fifth of the teacher-level variance reduced by modelling the stated controls and individual predictor variable. The school-level variance is significantly reduced in some of the models: by as much as 90% reduction when modelling 'work in small groups' or incorporating opportunities to relate learning in mathematics to 'daily lives'. The largest proportion reduction in variance for those particular predictor variables is attributed to school-level of the model, possibly reflecting the school ethos or structure as much as the individual teacher choices on when, and how frequently, to use group work or to incorporate links to daily lives. A fairly consistent reduction of just over 10% in student-level variance is reported in each model. Reduction in DIC further supports the significance of the models that provide explanation of variance through either significant coefficients in response categories (t-value >2.58 for Sig. level 1%); details on proportion of variance explained is also presented in Appendix 5.4.1.

Learning environment

In the G8 models of mathematics achievement, one of the most significant explanatory factors is the student response to lecture-style presentations. This variable provides the highest positive effect size of $\Delta = 0.34$ when lecture-style is reported for about half the lessons, and only marginally reduced to $\Delta = 0.28$ when witnessed every or almost every lesson. On wider classroom organisation, significant explanation of variance in the models is only evidenced when working individually or working in small groups 'every or almost every lesson'. There is a positive association for 'work problems on our own', where a significant effect is reported ($\Delta = 0.36$), whereas a stronger negative effect is noted for 'working together in small groups' every or almost every lesson, evidenced by an effect of $\Delta = -0.50$. Neither of the other response categories for these two variables provides significant explanation of variance in the multivariate model of mathematics achievement.

Reform-based practice and discussion

The most significant explanatory variable in this cluster concerns frequency of opportunity to memorising formulas and procedures. All three responses provide consistently significant explanation of variance with effect sizes of $\Delta = 0.25\ to\ 0.27$.

Unlike the combination of memorising and explaining reported for G4, there is no added benefit from considering the two variables together for G8 responses. The multivariate model highlights a dominant effect from opportunities to memorise formulas and procedures, which is broadly unaffected by the inclusion of demands to explain answers. Explaining on its own offers limited explanation of variance in the model with a small effect size of 0.23 to 0.21 when witnessed in about half the lessons or more frequently; these coefficients are significant at the 95% level ($t = 2.4$).

Technology

The technology cluster provides a significant explanation of variance in the models of mathematics achievement concerning the use of ICT. Responses to using a calculator provide significant explanation of variance in the model with positive effect sizes of 0.31 to 0.34, the higher effect associated with response of 'about half the lessons'. The same cannot be said for use of a computer in mathematics lessons or indeed for using a computer for schoolwork in and out-of-school. Both of those variables are negatively associated with mathematics achievement scores across all bar one of the response categories, with a strong association identified in the model when a computer is used every lesson ($\Delta = -0.39$) or in about half the lessons ($\Delta = -0.34$). There is strong negative association with achievement scores when a computer is used in support of schoolwork every day ($\Delta = -0.72$); the only exception to this negative trend is when computers are used to support school work a few times a year, where there is a small but highly significant positive effect reported in the multivariate model ($\Delta = 0.14$).

Assessment and Feedback

Another experience of note relates to student's frequency of doing a quiz or test. The model highlights a strong negative association with achievement when students report this as a feature of their classroom practice in about half their lessons ($\Delta = -0.49$), or in every or almost every lesson ($\Delta = -0.63$). The latter is equivalent to those students being outperformed by nearly 75% of the control group who claim to never have a quiz or test as part of their mathematics lesson.

Opportunities to review homework in G8 mathematics lessons are positively

associated with achievement with a small but significant explanation of variance provided when experienced in up to half the lessons ($\Delta = 0.14$); any more frequently and the association loses significance.

Active learning and practical activities

A negative association is evidence when students report regular engagement with interpreting data in tables, charts or graphs. This effect only applies to a small proportion of respondents (10%), but nevertheless the effect size of –0.36 ranks that variable in the upper region of explanatory variables within models of G8 mathematics achievement. A small but significant positive association is attributed to the variable on solving complex problems, where students decide on their own procedures for solving multi-step problems. This effect is noted across all response categories for that variable in the multivariate model; a highly significant explanation is witnessed when this occurs in some lessons ($\Delta = 0.14$). This response supports student autonomy in learning and active learning within G8 Mathematics experiences.

5.5.3. Modelling all predictor and control variables (G8 mathematics)

The final model for consideration takes G8 mathematics achievement as the dependent variable and models all the control variables and predictor variables previously considered for this subset of $TIMSS_{2007}$ data. This is a natural extension of the clusters analysed section 5.5.2 where the multivariate responses are presented in Table 5.5-3.

The findings for this model are presented in Appendix 5.4.2. In summary, the key findings are presented below in Table 5.5-4, reporting positive and negative association with mathematics achievement as before.

Table 5.5-4: Summary of key findings in model of G8 mathematics achievement data with all control and predictor variables included

Response	Effect size	
	Negative association	Positive association
Memorise procedures etc.(never)		
Some lessons		0.28
About half the lessons		0.36
Every or almost every lesson		0.46
Review our homework (never)		
Some lessons		0.12
Have a quiz or test (never)		
Some lessons	-0.40	
About half the lessons	-0.57	
Every or almost every lesson	-0.60	
Work on own (never)		
Every or almost every lesson		0.38
Work in small group (never)		
Every or almost every lesson	-0.35	
Use a calculator in class (never)		
About half the lessons		0.30
Every or almost every lesson		0.31
Use computers in math lesson(never)		
About half the lessons	-0.23	
Computer in support of sch wk(none)		
A few times a year		0.13
Every day	-0.64	
Relate learning to daily life (never)		
Every or almost every lesson	-0.14	
Interpret data		
Every or almost every lesson	-0.35	

In secondary mathematics education the profile of association with achievement in terms of reform-based practice presents a different picture from that reported in G4

mathematics. Here 'memorising formulae and procedures' is significantly associated with mathematics achievement, with $\Delta = 0.28$ for memorising in some lessons rising to $\Delta = 0.46$ for memorising in every or almost every lesson. Around one fifth of the students are affected by that higher effect, which is equivalent to those students outperforming 68% of the control group. In contrast, none of the response categories on explaining answers provide significant explanation of variance in the combined model; noting G4 data was positively associated with explaining answers as documented in Table 5.3-3.

On learning environment, where students report frequency of working on own or in groups, both approaches provide significant association with mathematics achievement. In the combined model of G8 mathematics achievement a significant positive association is retained between working problems on my own and achievement ($\Delta = 0.38$) whereas a negative association of similar magnitude ($\Delta = -0.35$) is noted between working in a small group and G8 mathematics achievement. Those significant associations only affect the response category for 'every or almost every lesson'; this impacts 34% of students for working on their own, but fewer than ten per cent of the cohort for working in a small group. Response categories on lecture-style presentation do not offer any significant explanation of the variance in the combined model for G8 mathematics; almost two thirds of the cohort report 'listening to their teacher give a lecture-style presentation' in almost every lesson.

The role of ICT in support of mathematics class work presents a positive association with achievement for calculator use in half the lessons or more frequently, both categories reporting the equivalent of students exceeding the scores of 62% of the control group who never use calculators. Those response categories affect over fifty per cent of the G8 cohort. On the other hand computer use in class and for out-of-school support in mathematics education is generally negatively associated with achievement scores. The significant effects highlight computer use in class, when occurring in about half the lessons, as being negatively associated with achievement ($\Delta = -0.23$); and computer use in support of mathematics school work as negatively associated with achievement ($\Delta = -0.64$) when witnessed every day. However, the latter response category only adversely affects a small proportion of the population. A small but significant positive association is reported for computer use in support of school work a few times a year, where the association with achievement has an effect size of $\Delta = 0.13$.

Opportunities for students to engage in active learning and practical activities through solving complex problems or interpreting tables, charts and graphs provide limited explanation in the model of achievement. Although data handling was positively associated with G4 mathematics achievement across all three response categories, in G8 only one

response provides a significant contribution to the model, with a negative association between interpreting data (every or almost every lesson) and mathematics achievement ($\Delta = -0.35$). Any association between solving complex problems and achievement scores that was reported within the multivariate model (Table 5.5-3) has been lost.

The final grouping that provides significant explanation in the combined model of G8 mathematics achievement relates to assessment and feedback. Reviewing homework on 'some lessons' is positively associated with G8 mathematics achievement, where a small but significant effect of $\Delta = 0.12$ is computed. Other responses in this cluster are all negatively associated with achievement scores, where having a quiz or test at all is associated with effect sizes of $\Delta = -0.40$, -0.57, and -0.60, as the frequency runs from some lessons, through half the lessons, to every or almost every lesson. Over three quarters of students experience a quiz or test in some lessons, but with that reported effect size it is equivalent to those students being out performed by 66% of the control group. Fewer than ten per cent of students report having such an experience every lesson, but for those students this is the equivalent to having 73% of the control group exceeding their scores.

6. Alternative Analyses

As set out in the chapter on methodological issues, three alternative approaches are considered alongside the method used in Chapter 5. Each is a variant on how the plausible value methodology is pursued within secondary analysis of survey data. The primary method considered plausible value as the unit of analysis in the hierarchical model, nested within student, teacher and school levels. This decision was counter to standard recommendations from the research community but was used to explore the potential of modelling plausible value as the unit of analysis within the $TIMSS_{2007}$ data sets. Those within IEA argue that the only safe and acceptable practice in analyses of data based on plausible values is to replicate the analysis using each plausible value in turn and then to combine findings using Rubin's Rules as detailed in Chapter 3. This method will be referenced as Combined Plausible Values (Rubin). Clearly this approach generates additional work for the analyst and ways are often sought to reduce the workload by either using only one plausible value, or by using the average of the plausible values, as the dependent variable. These two alternatives are respectively referenced as Single Plausible Value (SPV) and Average Plausible Value (AVPV) in the models that follow. The average of the student's five plausible values has already been used in the initial exploratory data analysis and as such is a derived variable in the data set. This section will report on the use of those alternative approaches to analyse achievement data, starting with the average plausible value as the dependent variable in the multi-level model, proceeding through to use of a single plausible value, and finally to undertake analyses using all five plausible values that are then combined for reporting purposes. Findings from the G4 and G8 primary analyses identified a suitable sub-set of variables that will be presented in parallel with those alternatives to empirically investigate how they each match up to the Combined Plausible Values (CPV) model.

6.1. Grade 4

6.1.1. Modelling an average of plausible values (AVPV)

An arithmetic average of the five plausible values assigned to each student is used as the dependent variable in this model of student achievement. The hierarchical model for G4 science achievement takes the form:

$$ASSSCI\_AV_{ijk} = \beta_0 + v_{0k} + u_{0jk} + e_{0ijk}$$

where the average plausible value is taken as the achievement score in what is now a three-level structure. The lowest level in the model is student (*i*), nested within teacher (*j*), and school (*k*) levels of analysis. Note the change in nomenclature for each level of the hierarchy, with *i, j* and *k* being used in a three-level model; and also that the data is returned to a short format by removing all duplicate student-cases that were generated to accommodate plausible values as IDCASE. The error in student achievement is presented separately for each level of the model; errors assumed to be drawn from a Normal distribution with mean 0 and variance σ^2. Diagnostic checks for normality and homogeneity of variance were administered in the same way as outlined in previous sections to confirm assumptions were not being violated. Models of G4 achievement include background control variables as before and then all of the predictor variables are included to replicate final models reported in Chapter 5.2.5 and 5.3.2. An unconditional model is computed as before, providing a base level that permits calculation of proportion of variance explained by the control and predictor variables. That null model also provides the basis for measuring effect size where the denominator in effect size calculation is taken as student level variance from the respective null model.

6.1.2. Modelling a single plausible value (SPV)

Analyses based upon a single plausible value should provide unbiased results if the imputation model is correct, but will underestimate the error variance since the imputation measurement error cannot be included when only one of the five plausible values is accounted for. The expectation is therefore to have larger standard errors due to the missing variance component and to have unbiased estimates; those estimates will lack precision as we already know from the primary analysis that they are drawn from a posterior distribution that has quite a large variance.

The hierarchical model based on a single plausible value for mathematics achievement will have three levels to its hierarchy with student (*i*), nested within teacher (*j*),

and school (*k*). Any one of the five plausible values can be taken as the dependent variable in the model – I have adopted to use the third plausible value, giving the general form:

$$ASSSCI03_{ijk} = \beta_0 + v_{0k} + u_{0jk} + e_{0ijk}$$

Models of G4 achievement are computed with background control variables as before, and then with each of the predictor variables incorporated to have a model of all predictor and control variables for comparison with earlier results. The null model based on a single plausible value is computed and used to generate effect sizes and proportion variance explained as above.

6.1.3. Modelling combined plausible values using Rubin's Rules (CPV-Rubin)

This third alternative method of analysis combines the findings from each of the separate SPV models, extending section 6.1.2. Once the five sets of parameter estimates and standard errors are generated, the estimates are combined using Rubin's Rules for handling plausible values (as outlined in Chapter 3). It is important to acknowledge plausible values are not test scores and cannot be treated as such, given they are generated using multiple imputation. Plausible values are random draws from the distribution of scores that could reasonably be assigned to each individual. As such the plausible values contain random error variance components and are not optimal scores for individuals; the random error variance was explicitly cited in primary analyses, where plausible value was taken as unit-level. The total variance in the combined model (CPV-Rubin) is increased to account for imputation error as well as sampling error, partitioning variance into two components that addressed within-imputation and between-imputation variances. A weighting in favour of the between-imputation variance is applied, where that weighting is related to the number of imputations or plausible values being combined. However, this weighting appears to be rather arbitrary in the supporting literature (Raudenbush, 2004; Rubin, 1988; Schafer, 1999; Sinharry, Stern & Russell, 2001) and makes for conservative parameter estimates. The parameter estimates from each plausible value were combined to generate an overall estimate ($\overline{Q}$) and its standard error, plus other statistics as detailed in Table 6.1-1.

Five key elements from each model were stored, taking the parameter estimate and its standard error, the significance level, effect size and its error term. The table shows the key elements from the first plausible value (columns B to F as extracted from an Excel sheet) and the later columns with associated formulae for the within-imputation and between-imputation variance to compute standard errors in the model. A summary of each alternative model for G4 mathematics is presented in Table 6.1-2 .

Table 6.1-1: Combining estimates and standard errors from plausible value methodology (Explain and Memorising)

Response ASMMAT01-05	Coef.	S.E.	Sig	effect size	+/- error	Overall estimate (Q bar, $\bar{Q}$)	Within Imputation Variance (U bar, $\bar{U}$)	Between Imputation Variance (B)	Total Variance (T)	Overall SE	Sig	effect size	+/- error
								$((B1-\bar{Q})^2+(G1-\bar{Q})^2$					
						AVERAGE	AVERAGE	$+(L1-\bar{Q})^2+(Q1-\bar{Q})^2$					
A	B	C	D	E	F	(B1,G1,L1,	($C1^2$,$H1^2$,	$+(V1-\bar{Q})^2)/4$	$\bar{U}+\left(1+\frac{1}{5}\right)B$				
						Q1,V1)	$M1^2$,$R1^2$,	Or					
							$W1^2$)	VAR.S					
								(B1,G1,L1,Q1,V1)					
EXPLAIN													
SOME LESSONS	9.6	3.6	**	0.14	0.05	9.2	13.1	1.04	14.32	3.8	*	0.13	0.06
ABOUT HALF THE LESSONS	13.2	3.9	**	0.19	0.06	13.5	15.2	0.38	15.65	4.0	**	0.20	0.06
EVERY OR ALMOST EVERY LESSON	10.9	4.1	**	0.16	0.06	11.2	16.7	0.95	17.81	4.2	**	0.16	0.06
MEMORISE													
SOME LESSONS	1.1	4.8	-	0.02	0.07	3.4	22.5	4.80	28.27	5.3	-	0.05	0.08
ABOUT HALF THE LESSONS	6.5	4.8	-	0.09	0.07	7.2	22.9	4.74	28.56	5.3	-	0.10	0.08
EVERY OR ALMOST EVERY LESSON	17.4	4.9	**	0.25	0.07	18.7	23.3	2.42	26.17	5.1	**	0.27	0.07

Table 6.1-2: Comparing Alternative Models for G4 Mathematics

	MLPV (ASMMATPV)		AVPV (ASMMAT_AV)		SPV (ASMMATO3)		CPV (ASMMAT_RUBIN)	
	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size
EXPLAIN								
SOME LESSONS	9.5 (3.3)**	0.15	9.5 (3.3)**	0.14	7.9 (3.6)*	0.12	9.5 (3.7)*	0.14
ABOUT HALF THE LESSONS	14.0 (3.6)**	0.22	14.0 (3.6)**	0.20	13.2 (3.9)**	0.19	14.0 (3.9)**	0.20
EVERY OR ALMOST EVERY LESS.	11.0 (3.7)**	0.17	11.0 (3.7)**	0.16	10.2 (4.1)*	0.15	11.1 (4.2)**	0.16
MEMORISE								
SOME LESSONS	3.8 (4.3)-	0.06	3.8 (4.3)-	0.06	6.1 (4.8)-	0.09	3.8 (5.3)-	0.06
ABOUT HALF THE LESSONS	7.3 (4.4)-	0.11	7.3 (4.4)-	0.11	8.3 (4.8)-	0.12	7.3 (5.3)-	0.11
EVERY OR ALMOST EVERY LESS.	19.1 (4.4)**	0.30	19.1 (4.4)**	0.28	19.1 (4.8)**	0.28	19.0 (5.1)**	0.28
MEASURE								
SOME LESSONS	1.9 (3.0)-	0.03	1.9 (3.0)-	0.03	2.3 (3.3)-	0.03	1.8 (3.5)-	0.03
ABOUT HALF THE LESSONS	-16.8 (4.2)**	-0.26	-16.8 (4.2)**	-0.24	-15.8 (4.6)**	-0.23	-16.7 (5.8)**	-0.24
EVERY OR ALMOST EVERY LESS.	-28.1 (4.7)**	-0.44	-28.1 (4.7)**	-0.41	-24.1 (5.2)**	-0.35	-28.1 (6.0)**	-0.41
TABLES, CHARTS & GRAPHS								
SOME LESSONS	24.9 (4.0)**	0.39	24.9 (4.0)**	0.36	22.7 (4.4)**	0.33	25.0 (4.7)**	0.37
ABOUT HALF THE LESSONS	20.1 (4.3)**	0.32	20.1 (4.3)**	0.29	19.1 (4.7)**	0.28	20.2 (5.0)**	0.30
EVERY OR ALMOST EVERY LESS.	13.8 (4.7)**	0.22	13.8 (4.7)**	0.20	10.8 (5.2)*	0.16	13.8 (6.0)**	0.20

Table 7.1-2: Comparing Alternative Models for G4 Mathematics (continued)

	MLPV (ASMMATPV)		AVPV (ASMMAT_AV)		SPV (ASMMATO3)		CPV (ASMMAT_RUBIN)	
	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size
WORK ON OWN								
SOME LESSONS	24.6 (7.5)**	0.39	24.6 (7.5)**	0.36	27.9 (8.3)**	0.41	25.2 (8.7)**	0.37
ABOUT HALF THE LESSONS	29.8 (7.5)**	0.47	29.8 (7.5)**	0.43	29.8 (8.3)**	0.43	30.3 (8.2)**	0.44
EVERY OR ALMOST EVERY LESS.	41.9 (7.4)**	0.66	41.9 (7.4)**	0.60	43.3 (8.1)**	0.63	42.5 (8.2)**	0.62
WORK IN SMALL GROUP								
SOME LESSONS	11.0 (4.1)**	0.17	11.0 (4.1)**	0.16	12.8 (4.5)**	0.19	11.1 (5.1)*	0.16
ABOUT HALF THE LESSONS	14.8 (4.5)**	0.23	14.8 (4.5)**	0.21	16.3 (4.9)**	0.24	14.8 (5.7)**	0.22
EVERY OR ALMOST EVERY LESS.	-1.2 (4.5)-	-0.02	-1.2 (4.5)-	-0.02	-0.6 (4.9)-	-0.01	-1.4 (5.8)-	-0.02
USE A CALCULATOR IN CLASS								
SOME LESSONS	12.4 (2.3)**	0.20	12.4 (2.3)**	0.18	13.0 (2.5)**	0.19	12.6 (2.6)**	0.18
ABOUT HALF THE LESSONS	4.4 (4.5)-	0.07	4.4 (4.5)-	0.06	8.0 (4.9)-	0.12	4.4 (5.7)-	0.06
EVERY OR ALMOST EVERY LESS.	-21.7 (7.4)**	-0.34	-21.7 (7.4)**	-0.31	-21.4 (8.1)**	-0.31	-21.8 (9.3)*	-0.32
USE A COMPUTER IN CLASS								
SOME LESSONS	-3.1 (2.3)-	-0.05	-3.1 (2.3)-	-0.04	-5.3 (2.5)*	-0.08	-3.2 (3.2)-	-0.05
ABOUT HALF THE LESSONS	-13.0 (3.6)**	-0.20	-13.0 (3.6)**	-0.19	-13.6 (4.0)**	-0.20	-12.9 (4.1)**	-0.19
EVERY OR ALMOST EVERY LESS.	-29.9 (4.4)**	-0.47	-29.9 (4.4)**	-0.43	-35.8 (4.8)**	-0.52	-30.0 (6.9)**	-0.44

Note: ** t-value>2.58, sig. level 1%; * t-value>1.96, sig. level 5%

256

6.1.4. Comparison of alternative models for G4 mathematics and science

Three main points emerge from these comparisons. First, the parameter estimates and standard errors for the MLPV and AVPV responses reported in Table 6.1-2 are identical. There are therefore no changes to significance levels between those two models on any of the parameters but reported effect sizes show minimal change from those computed earlier; the absolute value of effect size being slightly lower for tabulated coefficients in the AVPV model. This change can be attributed to the fact that the student-level variance component is larger as it incorporates the error previously assigned to the plausible value at unit-level of MLPV model. This increase in student-level variance will reduce effect size, given:

$$\text{Effect size} = \frac{Coefficient}{Student-level\ Variance}$$

Second where the parameter estimates in the SPV model are contrasted with results from the MLPV model, i.e. the primary model reported earlier. There is a high degree of fluctuation in the point estimates, ranging ± 5 points from values reported in the primary analysis, and a consistent increase of around ten per cent on standard errors of the estimates based on a single plausible value. Given those statistics are based on imputed data, with only a proportion of the data actually known to come from the individual respondent, the reported standard error is itself under-estimated. Therefore although the standard error has already been inflated from that calculated in the other models (AVPV and MLPV), we know that the standard error should be larger still as result of the imputation process. The findings are also governed by the choice of plausible value selected, i.e. a different set of conclusions could result from using one of the other plausible values. Given the standard errors are all higher than those computed in primary analyses, there is correspondingly less significance in the findings where the point estimates were similar; four of the parameter estimates either drop from 1% to 5% significance level, or become non-significant entries in the model, and one gains significance where the estimate from SPV was higher (use a computer in class some lessons). The strength of effect size similarly changes in line with the above, where SPV model has on average absolute reductions on effect size of up to 0.07 standard deviation units.

Third, the model coefficients and significance levels from the analysis of CPV (Rubin) highlight there are substantive changes in significance levels, with four of the responses losing significance when compared with data taken from the primary analysis. The changes in significance are in predictor variables with small effects in the model of mathematics achievement so the main substantive conclusions are generally unaffected. The

use of Rubin's Rules for combining findings has made an average rise of twenty two per cent on the standard errors reported in MLPV model. The effect sizes show very little change with a reduction in strength of effect for estimates in the CPV model of around 0.03 standard deviations. The main findings remain much as before, albeit with marginally smaller effect sizes and reduced level of significance. Changes are noted in level of significance in Table 6.1-2 for significant predictor variables i.e. explain answers (some lessons), make tables, charts and graphs (every or almost every lesson), work in small group (some lessons), and use a calculator in class (every or almost every lesson). Table 6.1-3 provides the significant effect sizes from the CPV model with MLPV effects in parenthesis.

Table 6.1-3: Significant findings in CPV model of G4 mathematics (MLPV effects)

Response	Effect size	
	Negative association	Positive association
Explain answers (never)		
Some lessons		0.14 (0.15)
About half the lessons		0.20 (0.22)
Every or almost every lesson		0.16 (0.17)
Memorise procedures etc.(never)		
Every or almost every lesson		0.28 (0.30)
Measure things in class (never)		
About half the lessons	-0.24 (-0.26)	
Every or almost every lesson	-0.41 (-0.44)	
Make tables, charts & graphs (never)		
Some lessons		0.37 (0.39)
About half the lessons		0.30 (0.32)
Every or almost every lesson		0.20 (0.22)
Work on own (never)		
Some lessons		0.37 (0.39)
About half the lessons		0.44 (0.47)
Every or almost every lesson		0.62 (0.66)
Work in group (never)		
Some lessons		0.16 (0.17)
About half the lessons		0.22 (0.23)
Use a calculator in class (never)		
Some lessons		0.18 (0.20)
Every or almost every lesson	-0.32 (-0.34)	
Computer in class (none)		
About half the lessons	-0.19 (-0.20)	
Every or almost every lesson	-0.44 (-0.47)	

Empirically evaluating alternative approaches to analysis of the survey data has shown that strong substantive associations with achievement can be identified through basic methods (AVPV and MLPV), but variables with marginal association may be claimed to carry a greater weight than warranted because of underestimating the standard error

associated with their parameter estimates. The same patterns of association are noted for the models of Science achievement at G4 presented in Table 6.1-5.

The intention behind developing a new multilevel plausible value method (MLPV) was to consider whether the multilevel structure and partitioning of variance could provide a mechanism that was an improvement on using the average of plausible values (AVPV) as those under-estimated the standard errors, but less demanding than pursuing Rubin's CPV method. It appears not, because the standard errors in the MLPV model are identical and therefore also too small. Although the variance partition coefficient for the plausible value level of the MLPV model is known, it does not get factored into the standard error calculation. While using the multilevel analysis has suitably addressed issues concerning analysis of clustered data and has provided transparency over the impact of imputing data on this scale, with plausible values accounting for nearly a quarter (23%) of unexplained variance in the final G4 model of science achievement, it has not corrected the standard errors in the model. The proportion of unexplained variance attributed to plausible value level of the other MLPV models, as illustrated in Appendix 6.3, is: 18% in G4 mathematics; 19% in G8 science; and 11% in G8 mathematics. The variance partition coefficients in the CPV models mirror those identified in the MLPV primary analysis at school- and teacher-level of the model, but the student-level variance absorbs the variance in plausible values. For example in G4 mathematics models the variance partition coefficient rises from 0.70 in the 4-level structure to 0.89 in the 3-level model (CPV). Those data clearly emphasise the variability attributed to imputation in survey data and the potential impact of basing analyses on plausible value methodology.

Although basic descriptive analyses and major associations with educational performance can be identified without resorting to the combined plausible value approach, caution is needed when drawing conclusions on anything other than strong associations. As can be seen from the empirical analyses, the level of significance reduces when standard errors are computed to take account of the imputation process; the only valid way of pursuing this accurately is to follow Rubin's rules for the analysis of imputed data.

The comparable models of G4 Science achievement presented in Table 6.1-5 provide similar evidence in support of the argument for using Rubin's rules for those analyses, with consistently increased standard errors that account for imputation and plausible value methodology. In G4 science the standard errors are increased on average by forty three per cent. There are five responses that have changed significance levels, either reduced from 1% to 5% sig. level or dropping significance altogether. As above, changes in significance are attributed to increased standard errors as there is minimal change in the

point estimates that are within a of range ± 0.37, an average change of 0.02 units.

Although strength of evidence on individual responses has been reduced, there is very little change in the computed effect sizes with only -0.05 to $+0.03$ change on effects from those reported under the MLPV model. Substantive messages are broadly unchanged with significant conclusions left intact. The principal positive associations with science achievement in the G4 Science model (Table 6.1-4) are: memorising science facts; doing and watching experiments or investigations; and working on own. There is one remaining negative association with scientific achievement that was identified in earlier models i.e. using a computer (about half the lessons, and every or almost every lesson). Three response categories lost significance in the model when the combined plausible value was used as the dependent variable, namely: watch science experiments (at least once a week); plan experiments (a few times a year); and work on own (a few times a year).

Table 6.1-4: Significant findings in CPV model of G4 science (MLPV effects)

Response	Effect size	
	Negative association	Positive association
Memorise science facts(never)		
Once or twice a month		0.20 (0.22)
At least once a week		0.36 (0.41)
Watch science experiments (never)		
A few times a year		0.21 (0.24)
At least once a week	(-0.13)	
Plan science experiments (never)		
A few times a year		(0.10)
Do science experiments (never)		
A few times a year		0.19 (0.21)
Once or twice a month		0.23 (0.26)
At least once a week		0.16 (0.18)
Work on own (never)		
A few times a year		(0.15)
Once or twice a month		0.19 (0.21)
At least once a week		0.22 (0.25)
Computer in science lesson (never)		
About half the lessons	-0.13 (-0.14)	
Every or almost every lesson	-0.30 (-0.33)	

The higher standard errors in the CPV model reflect the fact that although we may appear to increase the amount of data used to make parameter estimates, there is no *new* data, but merely further use of imputed data that we know little about and therefore have to further increase the standard errors using Rubin's factor.

Table 6.1-5: Comparing Alternative Models for G4 Science

	MLPV (ASSSCIPV)		AVPV (ASSSCI_AV)		SPV (ASSSCIO3)		CPV (ASSSCI_RUBIN)	
Response	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size
EXPLAIN SCIENCE STUDIED (NEVER)								
A FEW TIMES A YEAR	-1.3 (3.2)-	-0.02	-1.3 (3.2)-	-0.02	-2.3 (3.7)-	-0.03	-1.2 (5.1)-	-0.02
ONCE OR TWICE A MONTH	0.0 (3.3)-	0.00	0.0 (3.3)-	0.00	-3.0 (3.8)-	-0.04	0.1 (5.0)-	0.00
AT LEAST ONCE A WEEK	-2.6 (3.6)-	-0.04	-2.6 (3.6)-	-0.04	-4.9 (4.1)-	-0.07	-2.4 (5.0)-	-0.04
MEMORISE SCIENCE FACTS (NEVER)								
A FEW TIMES A YEAR	4.5 (3.2)-	0.07	4.5 (3.2)-	0.07	3.4 (3.6)-	0.05	4.4 (4.0)-	0.06
ONCE OR TWICE A MONTH	13.6 (3.2)**	0.22	13.6 (3.2)**	0.22	11.7 (3.6)**	0.17	13.4 (4.7)**	0.20
AT LEAST ONCE A WEEK	24.7 (3.2)**	0.41	24.7 (3.2)**	0.40	23.5 (3.7)**	0.34	24.6 (4.0)**	0.36
WATCH SCIENCE EXPERIMENTS (NEVER)								
A FEW TIMES A YEAR	14.3 (3.3)**	0.24	14.3 (3.3)**	0.23	14.8 (3.7)**	0.22	14.3 (4.5)**	0.21
ONCE OR TWICE A MONTH	3.3 (3.4)-	0.06	3.3 (3.4)-	0.05	4.3 (3.9)-	0.06	3.5 (4.4)-	0.05
AT LEAST ONCE A WEEK	-8.0 (3.6)*	-0.13	-8.0 (3.6)*	-0.13	-8.3 (4.1)*	-0.12	-7.8 (4.9)-	-0.11
PLAN SCIENCE EXPERIMENTS (NEVER)								
A FEW TIMES A YEAR	6.3 (2.7)*	0.10	6.3 (2.7)*	0.10	4.7 (3.0)-	0.07	6.3 (3.5)-	0.09
ONCE OR TWICE A MONTH	0.6 (3.0)-	0.01	0.6 (3.0)-	0.01	2.3 (3.4)-	0.03	0.6 (4.3)-	0.01
AT LEAST ONCE A WEEK	-0.7 (3.7)-	-0.01	-0.7 (3.7)-	-0.01	-3.9 (4.2)-	-0.06	-0.6 (5.5)-	-0.01
DO SCIENCE EXPERIMENTS (NEVER)								
A FEW TIMES A YEAR	12.5 (3.1)**	0.21	12.5 (3.1)**	0.20	12.5 (3.5)**	0.18	12.6 (4.9)**	0.19
ONCE OR TWICE A MONTH	15.6 (3.4)**	0.26	15.6 (3.4)**	0.25	17.4 (3.8)**	0.25	15.7 (4.9)**	0.23
AT LEAST ONCE A WEEK	10.7 (3.8)**	0.18	10.7 (3.8)**	0.17	12.4 (4.4)**	0.18	10.7 (5.3)*	0.16

Table 7.1 4: Comparing Alternative Models for G4 Science (continued)

	MLPV (ASSSCIPV)		AVPV (ASSSCI_AV)		SPV (ASSSCIO3)		CPV (ASSSCI_RUBIN)	
Response	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size
WORK ON OWN(NEVER)								
A FEW TIMES A YEAR	9.2 (4.0)*	0.15	9.2 (4.0)*	0.15	7.1 (4.5)-	0.10	9.1 (4.8)-	0.13
ONCE OR TWICE A MONTH	12.6 (3.7)**	0.21	12.6 (3.7)**	0.20	11.6 (4.2)**	0.17	12.6 (4.7)**	0.19
AT LEAST ONCE A WEEK	15.1 (3.6)**	0.25	15.1 (3.6)**	0.24	15.0 (4.1)**	0.22	15.1 (4.8)**	0.22
WORK IN GROUP (NEVER)								
A FEW TIMES A YEAR	2.8 (3.8)-	0.05	2.8 (3.8)-	0.05	1.7 (4.3)-	0.02	2.9 (7.5)-	0.04
ONCE OR TWICE A MONTH	0.2 (3.8)-	0.00	0.2 (3.8)-	0.00	-1.4 (4.3)-	-0.02	0.5 (6.4)-	0.01
AT LEAST ONCE A WEEK	-5.3 (4.0)-	-0.09	-5.3 (4.0)-	-0.09	-6.0 (4.5)-	-0.09	-5.0 (7.3)-	-0.07
COMPUTER USE (NEVER)								
SOME LESSONS	0.5 (2.6)-	0.01	0.5 (2.6)-	0.01	-0.4 (2.9)-	-0.01	0.3 (3.0)-	0.00
ABOUT HALF THE LESSONS	-8.5 (2.9)**	-0.14	-8.5 (2.9)**	-0.14	-7.6 (3.2)*	-0.11	-8.7 (3.7)*	-0.13
EVERY OR ALMOST EVERY LESS.	-20.2 (3.6)**	-0.33	-20.2 (3.6)**	-0.33	-21.1 (4.1)**	-0.31	-20.5 (5.2)**	-0.30

Note: ** t-value>2.58, sig. level 1%; * t-value>1.96, sig. level 5%

6.2. Grade 8

6.2.1. Alternative models for G8 mathematics and science

Similar findings are reported for the analyses of G8 data. The parameter estimates and standard errors for the MLPV and AVPV models of G8 Mathematics achievement scores are identical; similarly for the models of G8 science achievement. For the same reasons as cited in 6.1.4, only marginal changes of around 0.01 are noted on effect sizes within both models. The SPV models show volatility in point estimates and increased standard error. Those data not only emphasise the spread of responses across the models when one plausible value is used, but also illustrate the significant difference in findings that are dependent on the choice of plausible value. To illustrate this two entries are presented for science in Table 6.2-1, using SPV01 and SPV03 to show the differences between them in terms of point estimates and effect sizes.

Table 6.2-1: Summary of differences within estimates in analyses using SPV model

Dependent Variable	Range of change in coefficient estimates	Standard error Average Percentage change (%)	Range of change in effect size
Mathematics SPV			
BSMMAT03	-6.4 to +2.8	15	-0.19 to +0.12
Science SPV			
BSSSCI03	-8.5 to +2.8	10	-0.15 to +0.05
BSSSCI01	-11.7 to +5.6	11	-0.16 to +0.10

The third model uses the combined PV (Rubin) as the dependent variable, as outlined in Chapter 3 and illustrated in 6.1.3. Summarised data on each of the alternative models is provided in Appendix 6.2.1, but as above, summaries of differences within estimates from Rubin's model are highlighted in Table 6.2-2.

Table 6.2-2: Summary of differences within estimates in analyses using CPV model

Dependent Variable	Range of change in coefficient estimates	Standard error Average Percentage change (%)	Range of change in effect size
Mathematics CPV			
BSMMAT_RUBIN	-0.77 to +0.84	46	-0.05 to +0.08
Science CPV			
BSSSCI_RUBIN	-0.44 to +0.29	34	-0.03 to +0.02

As with the G4 data I conclude that the G8 CPV models are the only correct approaches as they incorporate appropriate standard errors to account for the imputation process. The MLPV and AVPV models both underestimate standard errors; the SPV model has inflated standard errors, but even those are under-estimated in relation to the recognised methodology following Rubin's rules for analysis of imputed data. The predictor variables with smaller effects are the ones where responses lose significance; an exception to this is 'use of a calculator in class' where the significance level is lost in all response categories, but that variable had, and indeed still retains, a strong association with G8 mathematics achievement scores ($\Delta = 0.27$ for half the lessons; $\Delta = 0.28$ for every or almost every lesson). This illustration highlights the benefit of reporting effect size alongside levels of significance that are computed using a t-test. Coe (2002) warns against such a singular reliance on p-values, noting that significance could be warranted either if the effect is big, or if the sample is very big. The p-value is conditional on the size of the effect *and* the size of the sample, so even though the effect size might be very small a large sample could support a significant result. Reporting effect size will turn attention to the strength of association with achievement, the impact of a particular response, and the size of its effect.

Science

The main findings stay much as identified under MLPV model. Of the seven changes in significance noted in the CPV model of G8 science achievement, three responses are no longer significant and the other four are now only significant at the 5% level. Details of comparisons between alternative models are provided in Appendix 6.2.1 with a summary of significant effect sizes from the CPV model presented in Table 6.2-3; MLPV effects are shown in parenthesis for direct comparison of changes due to using alternate model.

Table 6.2-3: Significant findings in CPV model of G8 science achievement (MLPV effects)

Response	Effect size	
	Negative association	Positive association
Explain science studied(never)		
Every or almost every lesson		0.20 (0.22)
Lecture-style presentation (never)		
Every or almost every lesson		(0.25)
Work problems on our own (never)		
Every or almost every lesson		0.17 (0.19)
Review our homework (never)		
About half the lessons	-0.20 (-0.23)	
Every or almost every lesson	-0.25 (-0.28)	
Have a quiz or test (never)		
Every or almost every lesson	(-0.25)	

Table 6.2 3: Significant findings in CPV model of G8 (Continued)

Response	Effect size	
	Negative association	Positive association
Use computers in sci lesson(never)		
About half the lessons	-0.17 (-0.18)	
Every or almost every lesson	-0.33 (-0.36)	
Computer in support of schwk(none)		
A few times a year		0.14 (0.16)
At least once a week	-0.16 (-0.17)	
Every day	(-0.28)	

Although the response on explaining science studied lost some significance, it still provides the strongest positive effect ($\Delta = 0.20$) in the CPV model. The only other significant positive associations are within work problems on own, and in occasional use of computer in support of school work (a few times a year). The significant negative associations are as reported earlier, with the exception of quiz or test that lost its significance within assessment and feedback; the other element in that strand on reviewing homework retains a highly significant and negative association with science achievement scores.

Mathematics

In the comparable model of mathematics achievement, six of the response categories are no longer significant in the CPV model of G8 mathematics achievement. Full details of comparisons are in Appendix 6.2.3 with summarised significant effects presented in Table 6.2-4. The principal associations that changed level of significance from those reported in the MLPV model are: review homework (some lessons); use a calculator in class (about half the lessons or more frequently); use a computer in class (about half the lessons); computer for school work (a few times a year); and daily life contexts (every or almost every lesson).

Prominent features of the learning environment and reform-based practice, including working on own, working in a group, and memorising processes, have retained a high level of significance in Rubin's model and support conclusions on pedagogical practices as outlined in Chapter 5.5.3. Similarly for practices related to assessment and feedback, where quiz or test has retained a strong negative association with mathematics achievement: one response category (some lessons) has a lower significance (5% level) than reported in the MLPV model but has retained a strong negative effect ($\Delta = -0.34$), equivalent to 63% of the control group exceeding the achievement score of students in that category. This rises to 69% of the control group when students report having quiz or test within every or almost every lesson ($\Delta = -0.51$).

Table 6.2-4: Significant findings in CPV model of G8 mathematics achievement (MLPV effects)

Response	Effect size	
	Negative association	Positive association
Memorise procedures etc.(never)		
Some lessons		0.25 (0.28)
About half the lessons		0.32 (0.36)
Every or almost every lesson		0.41 (0.46)
Review our homework (never)		
Some lessons		(0.12)
Have a quiz or test (never)		
Some lessons	-0.34 (-0.40)	
About half the lessons	-0.48 (-0.57)	
Every or almost every lesson	-0.51 (-0.60)	
Work on own (never)		
Every or almost every lesson		0.33 (0.38)
Work in small group (never)		
Every or almost every lesson	-0.30 (-0.35)	
Use a calculator in class (never)		
About half the lessons		(0.30)
Every or almost every lesson		(0.31)
Use computers in math lesson(never)		
About half the lessons	(-0.23)	
Computer in support of sch wk(none)		
A few times a year		(0.13)
Every day	-0.57 (-0.64)	
Relate learning to daily life (never)		
Every or almost every lesson	(-0.14)	
Interpret data		
Every or almost every lesson	-0.30 (-0.35)	

On inspecting the standard errors from the single plausible value method (SPV) it is already noted that they are under-estimated, but they are certainly closer to the appropriate parameter from the CPV model (Rubin) than those computed through AVPV or MLPV models. The SPV standard errors can be improved upon as additional plausible values are considered i.e. two plausible values provide a better estimate for the standard errors, three improves further and so on until all five are used following Rubin (1987). What may be worth exploring is whether a combination of AVPV and SPV models could be used as an alternative to the combined plausible value method; using AVPV to determine accurate point estimates, and the SPV model to generate a more realistic standard error, acknowledging that it is under estimated but not to the same degree as would be the case if AVPV model was used on its own. Clearly if one was interested in partitioning of variance then MLPV model could be substituted for AVPV. This could be investigated algebraically or by simulation to determine impact if any on final analyses of substantive issues.

7. Discussion and Conclusions

This study set out with a view to address two broad aspects of the $TIMSS_{2007}$ survey data, first to identify principal associations with students' educational performance and second to present a rationale and supporting evidence for an innovative approach to the analysis of survey data based on plausible values. In addressing these two aspects I sought to make the data more accessible, usable and applicable to practice, so that secondary data analysis of rich data sets such as TIMSS, PISA and other national survey data including SSLN could be pursued by a wider group of stakeholders. In addressing the first research question the analysis focused on two sets of variables. First there were aspects of the data that provided background detail on students, their homes, interests and personal characteristics that all served to provide explanation of variance within the data. Having controlled for variance in the target population, substantive issues of learning and teaching were analysed to identify principal features of practice and students' experiences that are significantly associated with educational achievements. This chapter will first consider those control variables that were used to control the data and will then proceed to evaluate the overall findings in relation to substantive, malleable features of learning and teaching that can provide evidence for policy and practice in schools. The last section of the chapter will draw conclusions on the methods of analysis that have been deployed, evaluating the merits of multilevel models of student achievement where first, plausible value is taken as the unit of analysis in a long data set with five records for each student (MLPV); second, the average of five plausible values is entered as the dependent variable in the model (AVPV); third, a single plausible value is used as the dependent variable (SPV); and finally comparing findings of each alternative method with those generated through the current gold standard in plausible value methodology, using all five plausible values in turn and combining parameter estimates using Rubin's rules for the analysis of imputed data (CPV). The empirical data provides evidence of similarities and differences between methods and illustrates why Rubin's method should be used in preference to the alternatives under consideration where precision and accuracy are essential elements of the analysis, to ensure conclusions are suitably robust by computing standard errors that reflect the nature of the data and the fact that it is an analysis of imputed values.

7.1. Background variables

The literature review of mathematics and science education centred on reform-based practice, with an emphasis on conceptual understanding and constructivism as a grand theory

of learning that would enhance students' development. Conceptual understanding is pursued for students to secure deep learning over short-term benefits that might be accrued from rote learning or instrumental understanding. A common thread of practice that lies within those theories of learning is talk and discussion, with active learning, practical experiments, and argumentation making the learning process very much a social activity. With that focus in mind it was useful to investigate and develop background variables that aligned themselves with some of those practices, for instance the place of talk and social activity that students engage in outwith their school setting, or knowing what resources or facilities are available to support learning at home. Although all elements of learning and background information cannot directly be accessed in the data there are useful indicators and associated practices that can be evaluated in relation to achievement scores in mathematics and science. How well learners achieve in mathematics and science will be influenced to some extent by their home background, personal characteristics and interests outwith any formal learning environment. Features of students' lives and out-of-school experiences can be controlled in the analysis to explain variance in the model and to condition the data set for a clearer analysis of the substantive issues under review. The classroom experiences and learning approaches espoused in the literature can be evaluated and studied in light of background characteristics that serve as proxy measures for social, economic and academic support in society. The set of home and individual variables identified for this purpose are as set out in Chapter 4.

Grade 4

The models of mathematics and science achievement scores generate parameter estimates, standard errors and effect sizes for each of the background variables submitted. The principal background characteristics were identified using those statistics to prioritise inclusion as control variables in models of achievement scores. Variables that provide evidence of a significant association with achievement scores will offer an explanation of variance in student achievement. Those associations may be quite different for each discipline, so this discussion will first consider features that are pertinent to both disciplines within the grade and second will focus on important differences in the parameter estimates that can be taken on board by practitioner and policy communities. In the discussion, links are made to literature to furnish stakeholders with information that can then be used alongside the parameter statistics that come from the variable. Strength of evidence is also included, using effect sizes from final combined plausible value models to allow stakeholders to make informed decisions on the basis of the data and to act accordingly.

The similarities in the G4 data are that both disciplines have the following four variables identified as offering greatest explanation in the models:

- Being born outwith UK and arrived aged 5 or over: $\Delta_S = -0.60$ and $\Delta_M = -0.53$
- Fewer books at home than reference category of 26 to 100 books; '0 to 10 books' is reported with effect sizes $\Delta_S = -0.42$ and $\Delta_M = -0.49$
- Never speaking English at home: $\Delta_S = -0.35$ and $\Delta_M = -0.28$
- Low perception of safety in school: $\Delta_S = -0.22$ and $\Delta_M = -0.17$

The effect sizes attributed to those key variables emphasise the relevance of including those elements as control variables when developing models of achievement scores. Schools need to be aware of negative associations with achievement to encourage appropriate actions to be taken, whether that relates to strengthening parental engagement with affected groups, or revisiting school policies on ethos and safety given low perceptions of safety are associated with lower achievements. Other background variables entered in the models appear to offer greater explanation in the Science data:

- More books at home than reference category of 26 to 100 books, with 'three or more bookcases (over 200 books)': $\Delta_S = 0.32$ and $\Delta_M = 0.22$; also with two bookcases (101 to 200 books): $\Delta_S = 0.24$ and $\Delta_M = 0.16$
- Home ICT with an internet connection: $\Delta_S = 0.39$ and $\Delta_M = 0.32$
- English language spoken at home 'sometimes': $\Delta_S = -0.29$ and $\Delta_M = -0.17$
- Not possessing a mobile phone: $\Delta_S = 0.27$ and $\Delta_M = 0.17$ (noting that *not possessing* a mobile phone has a *positive* effect in the models of achievement)

The level of home support, as reported through number of books at home and possession of an internet-connected computer, appears to have a greater effect in the model of science achievement scores. The number of books at home provides a strong and consistent association with achievement where significant variances are accounted for when modelled. Generally speaking the number of books at home is directly related to the achievement score, with any decrease/increase in the number of books being associated with lower/higher achievement scores relative to the reference response of 26-100 books. Both disciplines have features to be addressed in terms of supporting migrant students and their parents where English is not the first language in the household, noting that 'only sometimes speaking English at home' is a response that appears to have a stronger association with science rather than mathematics achievement scores. Home support and linguistic skills appear to be more important in developing scientific understanding and achievement.

Availability of a home computer with internet functionality offers a strong

association with achievement, providing leverage for improved broadband capability particularly in rural communities where greater benefits might ensue. A defining factor in future developments will be how computers are used in the broader context of home and school, particularly given findings on use of ICT in lessons, where a strong negative association was reported. Cuban's work on computers in the classroom and at home is still very valid today as he makes the point that computers and other technologies are used more at home than in the classroom (Cuban, 2001). Trying to get systemic change is difficult when there is no real consensus on what is needed or wanted in terms of computer literacy. He contends that even when there are committed and dedicated practitioners to champion ICT there is a tendency to stop short of real change in practice: 'although promoters of new technologies often spout the rhetoric of fundamental change, few have pursued deep and comprehensive change in the existing system of schooling' (Cuban, 2001:195).

Other variables are identified as providing greater explanation of variance in models of mathematics achievement, for example:

- Fewer books at home than reference category of 26 to 100 books, with 'one shelf (11 to 25 books)': $\Delta_S = -0.22$ and $\Delta_M = -0.28$
- Gender differences with girls under-achieving in relation to reference group of boys: $\Delta_S = -0.15$ and $\Delta_M = -0.30$

The gender differences are more pronounced in models of mathematics achievement than in the science models, but again, both disciplines have scope to redress such significant differences as reported in those models. There is no single major contributing factor behind the gender differences in these models of achievement that contain all predictor variables. However some variables play a slightly more influential role with the number of books at home, and out-of-school interests of a social dimension (play and talk with friends; play sports) heading the list from within the background variables. The predictor variables on ICT in school, using either a calculator or computer in class, appear to have the biggest individual influence on gender difference but as with the background variables no individual predictor provides a lot of explanation for observed gender differences. There is no clear consensus on how to explain gender differences in mathematics and science achievement scores with many authors better at eliminating claims rather than suggesting positive explanations (Brown and Brown, 2007; Fennema, 2002; Powney, 1996; Spelke, 2005). Biological and environmental based arguments rise and fall, with other features on learning styles and assessment, including the content of assessment and what is being assessed, unable to provide any consensually agreed solution to the concerns. Powney (1996: 63)

provides a summary of potential explanations, acknowledging there is no single factor but more the 'cumulative disadvantage accrued from a complex interaction of social and educational factors which affect a pupil's uptake and success in educational opportunities but it is not clear how boys and girls are influenced differently'. Geist & King (2008) set out ten things teachers can do to support mathematics for both genders; as well as stating the need to avoid labelling and pervasive gender stereotyping, their list includes several aspects of learning that are of a social nature, and that very much reside within reform-based practice. For example: using active and exploratory methods with opportunities for visual-spatial tasks as a contrast to language based mathematics; allowing children to solve things in different ways to promote thinking and discussion; cooperative learning and group work where learning mathematics is viewed as a social endeavour; all of which should support both genders in their pursuit of mathematics and science achievement.

In summary, and to address research question 1(a), the background characteristics of G4 students that are significantly associated with achievement scores can be classified under three headings of home support, migrant status and school practices. The principal characteristics of interest under home support are the number of books at home, the availability of study tools, and possession of a computer with internet connection. These serve as proxies for parental education and engagement, and for social and economic status of home environment. Background knowledge on migrant students that can explain variance in the achievement models, includes the age of arrival in UK and level of English that is spoken at home. Finally there are two aspects that pick up on schools' internal practices. The first of those is related to gender in STEM education, noting that girls are consistently associated with lower achievement scores in mathematics and science; and the second concerns students' perception of safety whilst in school. All of these features explain significant variance in models of achievement and as such are included as control variables before examining specific pedagogical practices and their association with mathematics or science achievement.

Grade 8

The background data variables from G8 present an analogous picture with similarities across disciplines, but also noteworthy differences dependent on whether it is mathematics or science achievement that is modelled. Any distinctions made between disciplines in the secondary sector are more relevant for schools, practitioners and other interested parties since the norm is for students to be taught by a different teacher for mathematics and science, or even to have multiple teachers in science; whereas more

generalist teachers work with students in the G4 cohort. There continues to be a gender difference in the models of mathematics and science achievement scores, but the strength of effect has changed for those older students: the association between gender and mathematics achievement has shown signs of weakening with a marginal reduction in effect size, whereas there is a strengthening of effect in science. The reported effect sizes are $\Delta_S = -0.25$ and $\Delta_M = -0.26$, equivalent to boys exceeding the achievement scores of 60 per cent of the girls' cohort. There is clearly a challenge for practitioners to take appropriate actions in both mathematics and science classrooms. The practices outlined earlier would serve well in addressing the gender challenge (Geist and King, 2008; Powney, 1996) and although a lot of gender literature is somewhat dated, the research and guidance is still valid with adopted strategies showing marginal gains over time; it is just a slow gain. Another source of instructional strategies that address gender differences and the reform agenda can be sourced from Morrow and Morrow (2005) where *Instructional Strategies* are identified by Kathie Anderson *et al.*

The level of home support provides a stronger association with G8 science achievement scores than for mathematics. This is much as reported for G4, but with the addition of possessing of own bedroom to the previously discussed books, study tools and home computer. The bedroom variable serves as a further proxy for SES as well as being an indicator of having a potential study space at home; those students without a study desk or space to study at home (around 15 per cent of target population) may need to be catered for through study clubs or having dedicated time for study in school resource areas or equivalent. The empirical data highlights lack of home support is negatively associated with achievement, leaving practitioners and school leaders to consider ways of redressing the impact of that background effect and its association with achievement scores.

The data on migrant students in Grade 8 takes account of older arrivals to the UK for this cohort of thirteen to fourteen year olds. Although the effect size has reduced for the equivalent group of new arrivals (aged over ten as opposed to over five years) where $\Delta_S = -0.49$ and $\Delta_M = -0.35$ [compared to $\Delta_S = -0.60$ and $\Delta_M = -0.53$ in G4], there remains a strong negative association in relation to 'never using English' at home. That effect in science rises to $\Delta_S = -0.57$ for the small group of approximately 700 in the target population who never speak English at home. As with G4 findings this factor has a weaker, association with the students' mathematics achievement scores; indeed in G8 there is a non-significant association, with $\Delta_M = -0.34$.

In summary, and to address the second part of research question 1(a), the background characteristics of G8 students that are strongly associated with achievement

scores can be categorised under the same three headings of home support, migrant status and school practices. The principal characteristics of interest under home support are those listed in G4 plus knowledge on possession of own bedroom/ study space when considering models of science achievement. Background knowledge on migrant students' use of English at home is particularly influential in science education; age of arrival in UK is a significant contributor in both disciplines for students arriving in UK older than 10. On school practices there is scope to review policy and practice in relation to gender issues and to consider ways of supporting study skills for those students in homes without a study space (own bedroom) or who are working in paid employment that may encroach on study time. The latter may be addressed by bringing forward to the earlier years of secondary school any existing practices on study skills that are currently targeted at students preparing for national qualifications.

7.2. Substantive issues

The conceptual framework developed in Chapter 2 was used to provide a focus for the evaluation of practice, where frequency of exposure to various experiences was examined for association with achievement scores as reported through $TIMSS_{2007}$. An important finding in this analysis is that a number of research and policy driven entries within the framework of malleable substantive practices do not explain variance in the models. Other practices that are significantly associated with achievement scores are not always in line with anticipated direction of association, highlighting negative associations where literature and policy initiatives allude to positive benefits from those practices. Table 7.2-1 and Table 7.2-2 provide information on those associations for G4 and G8 respectively, with non-significant practices scored out to highlight experiences that failed to provide evidence of association in the final models of mathematics and science achievement scores.

Table 7.2-1: Predictors that provide a significantly $+^{ve}$ or $-^{ve}$ association with G4 achievement; strike through font indicates factors that are not significant

How often do you do these things in your … mathematics lessons (G4)		science lessons (G4)	
Reform-based practice and discussion			
Explain my answers	$(+^{ve})$	~~Write or give an explanation for something I am studying in science~~	
Memorise how to work problems	$(+^{ve})$	Memorising science facts and procedures	$(+^{ve})$
Active learning and practical activities			
Measure things in the classroom and around the school	$(-^{ve})$	Watch the teacher do a science experiment	$(+^{ve})$
Make tables, charts or graphs	$(+^{ve})$	~~Design or plan a science experiment or investigation~~	
		Do a science experiment or investigation	$(+^{ve})$

Table 7.2-1: Predictors that are significantly associated with G4 achievement (Continued)

How often do you do these things in your … mathematics lessons (G4)		science lessons (G4)	
Learning environment			
Work problems on my own	(+ve)	Work science problems on my own	(+ve)
Work with other students in small groups	(+ve)	~~Work with other students in a small group on a science experiment or investigation~~	
Technology			
Use a calculator	(+ve)	Use a computer in science lessons	(−ve)
We use computers	(−ve)	Use a computer for science schoolwork (in and out of school)	(−ve)
Use a computer for mathematics schoolwork (in and out of school)	(−ve)		

Table 7.2-2: Predictors that provide a significantly +ve or −ve association with G8 achievement; strike through font indicates factors that are not significant

How often do you do these things in your … mathematics lessons (G8)		science lessons (G8)	
Reform-based practice and discussion			
~~Explain our answers~~		Give explanations about what we are studying	(+ve)
Memorise formulas and procedures	(+ve)	~~Memorise science facts and principles~~	
Active learning and practical activities			
Interpret data in tables, charts or graphs	(−ve)	~~Watch the teacher demonstrate an experiment or investigation~~	
~~Decide on our own procedures for solving complex problems~~		~~Design or plan an experiment or investigation~~	
		~~Conduct an experiment or investigation~~	
Learning environment			
Work problems on our own	(+ve)	Work problems on our own	(−ve)
Work together in small groups	(−ve)	~~Work in small groups on an experiment or investigation~~	
~~Listen to the teacher give a lecture-style presentation~~		~~Listen to the teacher give a lecture-style presentation~~	
Contexts (real life)			
~~Relate what we are learning in mathematics to our daily life~~		~~Relate what is learned in science to our daily lives~~	
Technology			
~~We use calculators~~		We use computers	(−ve)
~~We use computers~~		Use a computer for science schoolwork (in and out of school) ‘a few times a year’	(+ve)
Use a computer for mathematics schoolwork (in and out of school)	(−ve)	‘at least once a week’	(−ve)
Assessment and feedback			
~~Review our homework~~		Review our homework	(−ve)
Have a quiz or test	(−ve)	~~Have a quiz or test~~	

The final models presented at the end of each sub-section of Chapter 5 combine all predictor and control variables previously modelled for mathematics and science achievement. Those models are updated in line with combined plausible value (CPV) models and used to evaluate substantive issues in learning and teaching. For reporting purposes in this section only those variables that retained significance through the CPV model are included; the effect sizes for each response are as presented in Chapter 6.

7.2.1. Mathematics

All G4 mathematics predictor variables appear to retain significance within the CPV model, with a balanced experience of different practices showing a positive association with achievement scores. The most significant predictor reports on opportunities to work on your own, where all three responses are highly significant contributors to the model with the effect size providing evidence of strengthening association with achievement scores as the frequency of opportunity increases. The reference group of 'never' is understandably quite small so it makes sense to evaluate the change in effect across the three other response categories; those effects are incrementally larger as the frequency of working on own increases, progressing from $\Delta = 0.37$ through $\Delta = 0.44$ to $\Delta = 0.62$. A second variable on the learning environment considers the opportunity students have to work in small groups. As noted above the respondents indicate a balanced approach with both individual and group work providing significant positive associations with mathematics achievement scores. The associated effects for group work are smaller but nonetheless significant in the overall model with $\Delta = 0.22$ for working in small groups in about half the lessons; $\Delta = 0.16$ for 'some' lessons, which was the dominant response category with just under fifty per cent of students indicating that frequency of group work.

The active and practical dimension of making tables, charts or graphs provides the next most significant effect in the overall model of G4 achievement, with $\Delta = 0.37$ for some lessons being used to further this practical aspect of the curriculum. Over fifty per cent of respondents are in that category, with a further quarter in the neighbouring response of about half the lessons where effect size was $\Delta = 0.30$. Opportunities to engage with data handling provide a positive association with mathematics achievement scores. In contrast, the other practical experience in the model that reports on students measuring things in G4 class is negatively associated with achievement. Just under a fifth of students report this experience in more than 'some' lessons; response categories show a strong negative association with mathematics achievement through effect sizes of $\Delta = -0.24$ and $\Delta = -0.41$ as the frequency of experience increases to every or almost every day.

The G4 model provides a positive association with achievement in relation to the reform-based practice debate and the place of discussion in learning mathematics. This is where the data again highlights a balanced approach when evaluating effects associated with mathematics achievement scores, combining opportunities to explain and memorise. All three response categories on frequency of opportunity to 'explain answers' provide significant and positive effects, ranging from $\Delta = 0.14$ to $\Delta = 0.20$. The highest effect occurs when explaining is witnessed in about half the lessons, an experience that almost a quarter of the target population report as a feature of their learning. Memorising formulae and procedures on the other hand only provides a significant level of explanation when witnessed in every, or almost every lesson, a position held by around a third of the cohort where an effect size of $\Delta = 0.28$ is reported. A much smaller non-significant effect of $\Delta = 0.11$ is noted when memorising is a feature in about half the lessons. An eighth of the cohort never explains their answers, almost double the percentage of respondents who report never memorising.

Use of technology in learning is not well supported in the models, with the only positive effect being for occasional use of a calculator in the G4 model where $\Delta = 0.18$; when using a calculator extends to more frequent use the effect is reversed and strengthened with $\Delta = -0.32$ (every or almost every lesson). All other references to calculator or computer use in class show a negative association with mathematics achievement scores. Use of a computer in mathematics classes has a negative association with achievement, with effect size of $\Delta = -0.19$ when computer use is witnessed in about half the lessons, extending to $\Delta = -0.44$ when computers are used every or almost every lesson.

Fewer entries from the G8 model provide evidence of association with mathematics achievement. Working on your own (every or nearly every lesson) is positively associated with achievement scores ($\Delta = 0.33$), but the same cannot be reported for group work where a significant negative association is recorded. No level of group work at G8 offers a positive effect in the model of mathematics achievement, with an effect of $\Delta = -0.30$ for working in a group every lesson. This finding is counter to claims made in literature and policy documentation where it is advocated that collaborative learning enhances opportunities for understanding (Rogoff, 1994; Correa-Chavez and Rogoff 2005; Wenger 2006; HMIE, 2006).

At G8 the strongest association with mathematics achievement data is very clearly memorising formulae and procedures, where an effect size of $\Delta = 0.41$ is recorded for a response of every or almost every lesson; equivalent to exceeding the scores of 66% of the reference group (never memorising). Almost a fifth of the target population indicate that level of memorising in class but the other response categories also report large effects with

$\Delta = 0.32$ and $\Delta = 0.25$ for about half the lessons and some lessons respectively; all three effects are among the highest in the model. The association between explaining and achievement scores in G8 are all non-significant on the computed t-values; that was also the case for MLPV model so this loss of significance is not just down to increased standard errors resulting from Rubin's rules.

The only significant explanatory variable within the technology cluster at G8 is using a computer in support of school work every day, which has a strong negative association with achievement ($\Delta = -0.57$); all other references to calculator or computer use in class are non-significant in the CPV model of G8 mathematics achievement.

These findings provide some support for a reform agenda with benefits accrued from working together in G4. The dominant association with higher achievement in both G4 and G8 is a learning environment where students work individually, an experience that is witnessed in about half the lessons or more frequently by over three quarters of the target population. About fifty per cent of the cohort either never work in small groups or only do so in some lessons. The positive effect associated with group work in G4 could provide the impetus to strive for an increase in the proportion of students who experience group work and the associated collaboration, discussion and articulation. The aim might be to create opportunities for group work in about half the lessons because the effect weakens when increased to almost every lesson. Practitioners who do pursue group work on a more regular basis should take stock of the empirical findings that show the zero gain in the model from primarily pursuing group work; if nothing else, other teaching strategies and organisation of the learning environment should be considered in parallel.

If the drive for conceptual understanding is to be pursued then a shift to increase opportunities for students to explain their thinking and answers in tandem with memorising would be a useful step; the argument for such a move is supported by the G4 empirical data. Memorising is clearly a beneficial feature of learning, and undoubtedly a necessary component of learning, but for practitioners the emphasis must be on how that learning is progressed. Given the strength of association between explaining and mathematical achievement in the G4 data, the potential is there to have an impact on achievement in G8 if those practices are further developed. It is hard to be solely relying on memorising in a rote and instrumental fashion because there is too much to 'remember'; any shift towards a more balanced approach that encourages memorising *with* explaining and understanding should be beneficial.

Technology in schools is clearly not being fully exploited and these data further support the messages conveyed by Cuban (2001) where there is a need to consider how

technologies are best used in education. Big investments are currently being made with hardware, including the use of tablet technology in Scottish Primary schools. The Scottish Government has five objectives for taking forward the use of technologies in schools with the first being to change the culture in the use of ICT. This is especially relevant where access can be restricted and mobile devices are frowned upon or banned; effectively creating technology free zones rather technology rich environments that capitalise on the opportunities for communication, collaboration and extending learning beyond the confines of the individual educational establishment. A barrier in G8 that was highlighted by Macintyre & Forbes (2002) is the reluctance to embrace technologies that are not permitted in national examinations, even though those national qualification assessments may be several years ahead. Many opportunities to integrate technologies are acknowledged, but difficulties in separating learning from assessment appear to be the biggest barrier to further the use of Computer Algebra Systems (CAS) or other technologies that are not permitted under examination regulations. In parallel with a change of culture there will need to be professional development, not only to satisfy the second objective of improving confidence in the use of ICT but also to buy into that new culture with clarity of purpose – returning to Cuban's concern over achieving systemic change that is deep, comprehensive and sustainable (Cuban, 2001). Looking at historical developments of technological investments in education this is unlikely to be achieved through individual ambassadors; without a critical mass of confident and clear-minded practitioners it will be difficult to achieve systemic change.

One additional variable from the G8 analysis that is highlighted as being significantly associated with achievement scores, concerns the frequency and effect of having a quiz or test. The reported effects within this assessment and feedback variable are all negative for the response categories of some lessons, about half the lessons, through to every or almost every lesson, with respective effects of $\Delta = -0.34,\ -0.48$ and -0.51. This indicates the association with achievement scores is counter to the claims made on feedback and assessment within the AifL movement, and raises questions over how those instruments are being deployed if there is such a strong negative association with achievement scores. The other entry in that cluster, reviewing homework, fails to support the claim made by Hattie and Timperley (2007) on the value of feedback. This is a variable that showed marginal significance in the MLPV model but once the correct standard errors were computed for the imputed data the association with achievement was non-significant. Another claim in the literature that is left unsupported by the empirical data concerns opportunities to relate learning to daily life. At G4 this was modelled through a teacher

variable that was non-significant and at G8 the student responses are all non-significant in the CPV model; one response was marginally significant on the MLPV model, attaining significance at 5 per cent level with an associated effect of -0.14, but once the higher standard error was used this became non-significant with effect of $\Delta = -0.13$. These findings raise questions over the policy initiative and practice as implemented in schools. The intention is clearly to enhance learning and engagement through student ownership and acknowledgement of how the mathematics they are learning can be used in contexts, but this cannot be happening for it to have a negative association with achievement. There will be benefit in exploring the potential of using Realistic Mathematics Education (RME), the Dutch approach to mathematics education. In that movement the focus is not solely on real-world encounters, although it includes those, but it is more about being real for the learners with the derivation of RME stemming from the Dutch translation of 'to imagine', which is 'zich REALISEren' (Van den Heuvel-Panhuizen, M., 2005). This links into the seven principles of curriculum design espoused by Curriculum for Excellence, where all learners are to benefit from opportunities that embrace: Challenge and enjoyment; Breadth; Progression; Depth; Personalisation and choice; Coherence; and Relevance. A frequently cited sub-set of those principles related to assessment, calls for evidence of breadth, challenge and application in students' development. Practitioners could therefore draw on the RME research and well-supported techniques to incorporate use of contexts and daily life activities in support of students' ability to apply their learning to new situations.

In summarising the principal malleable experiences associated with achievement in mathematics, and to address research question 1 (b), Table 7.2-3 ranks associations by highest individual effect within the respective models and makes cross-reference to the conceptual framework set out in Chapter 2, i.e. Reform-based Practice and Discussion (RPD); Active Learning and Practical Activities (ALPA); Learning Environment (LE); Contexts (C); Technology (T); Assessment and Feedback (AF).

Table 7.2-3: Summary of Associations with Achievement in Mathematics (conceptual link and highest individual effect)

	Positive association	*Negative association*
Grade 4	Work on own (LE, Δ=0.62)	Use computers (T, Δ= -0.44)
	Make tables, charts and graphs (ALP, Δ=0.37)	Measuring in class (ALPA, Δ= -0.41)
	Memorise procedures (RPD, Δ=0.28)	Use a calculator (T, Δ= -0.32)
	Work in small group (LE, Δ=0.22)	
	Explain answers (RPD, Δ=0.20)	
	Use a calculator (T, Δ=0.18)	

Table 7.2 3: Summary of Associations with Achievement in Mathematics (continued)

	Positive association	*Negative association*
Grade 8	Memorise (RPD, Δ=0.41) Work on own (LE, Δ=0.33)	Use computer in support of school work (T, Δ= -0.57) Have a quiz or test (AF, Δ= -0.51) Interpret graphs, charts and tables (ALPA, Δ= -0.30) Work in a group (LE, Δ= -0.30)

Differences across stage within mathematics

Considering these findings in relation to the literature and policy initiatives discussed in Chapter 2 there are some notable differences in practice between G4 and G8 that can be reported to address research question 1 (d). First, taking the learning environment, we can see that opportunities for 'working on own' is strongly associated with both stages but more widely impacting on achievement in G4, where all three response options report a high effect size; there is only a significant association in G8 for every or almost every lesson, and even then the effect is almost half of that reported in G4. A more interesting and concerning finding is that opportunities for group work (every or almost every lesson) is negatively associated with mathematics achievement in G8, whereas the same learning environment in G4 is positively associated with achievement. Given the policy drive for collaborative and active learning this finding raises questions over the implementation and impact of collaborative activities in secondary schools, calling for more focused research on that practice.

Second, the use of calculator technology in classes only shows a positive association with achievement when used in some G4 lessons. Where there is more extensive use in primary school settings, rising to every or almost every lesson, a large negative association with achievement is reported. In contrast, there are no significant associations between calculator use and mathematics achievement in secondary schools. Given contemporary policy initiatives call for increased use of technologies in the widest sense, it is concerning that this finding highlights calculators are not being utilised or exploited to enhance learning and mathematical achievements. Similarly where it comes to computer technology in classrooms, where both G4 and G8 stages report strong negative associations between computer use and achievement in mathematics scores. Those findings appear to be about a culture change in the use of ICT, where there is a need to create and support an environment where technologies can be fully embraced and capitalised upon, and in addition to reap suggested benefits from related opportunities to discuss, collaborate and communicate

beyond the confines of host establishment.

A third area of difference relates to reform-based practice and discussion in schools. The experience of explaining how to work out problems is positively associated with achievement across all three response categories in G4 but does not provide any significant explanation of variance in G8 model of mathematics achievement. The dominant feature in G8 is memorising facts and procedures, something that also has a positive association in G4 but in what looks to be a more balanced experience in conjunction with explaining. Memorising clearly plays a part in learning but secondary students would appear to be missing out on reform-based practice of explaining through discussion and talk, a feature of learning that primary students appear to benefit from.

Finally, it is noticeable that there is a difference across stages for mathematics within the study of tables, charts and graphs (TCG). Given an increased focus on data handling and statistics within school mathematics and numeracy it is concerning that there is a swing from a positive association in G4 (where emphasis lies on making TCG) to a negative association with achievement in G8 (where emphasis shifts to interpretation of TCG). It is worth investigating what lies behind those findings to determine why students exposed to regular engagement with data are associated with lower achievement scores than those who are never called upon to interpret data.

7.2.2. Science

Memorising facts in science is very clearly a significant activity in G4 and as discussed in section 5.2.2, the joint exposure to memorising and explaining science studied can have an increased effect in the model. Explaining a few times a year, in conjunction with memorising at least once a week, looked as though it would offer a larger overall effect but in the final combined model of science achievement presented in 5.2.5 there is no evidence that explaining science studied contributes a significant explanation of variance in the model of G4 science achievement.

In terms of classroom organisation, two of the three responses to 'working on their own' in a science setting, provided significant explanation in the G4 model with effect sizes in the range of 0.19 to 0.22. Insofar as working in groups is concerned, the responses were not significant at the 5% significance level but for two of the responses (about half the lessons, and every or almost every lesson) the effect sizes were $\Delta = 0.21$ and $\Delta = 0.18$ respectively; this suggests there is an association with achievement scores but there is insufficient support on the current model to make any statistically significant claims.

A major area of interest in science education concerns the place of practical

experiments and investigations and their impact on measures of achievement. Evaluating effects for the cluster of variables linked to G4 science experiments or investigations (watch, do, and plan), it is evident that the most consistent impact comes through students 'doing' science experiments. When all predictors are modelled together, doing science experiments is the most significant contributor to the model with effects in the range $\Delta = 0.16$ to $\Delta = 0.23$, peaking when witnessed once or twice a month. Watching a teacher demonstration of an experiment offers a further positive effect of $\Delta = 0.21$ when witnessed a few times a year.

The role of ICT in learning and teaching does not appear to have any positive association with achievement in Science. Indeed for computer use in G4 there is a significant negative association with achievement, with $\Delta = -0.13$ (about half the lessons) and $\Delta = -0.30$ (for every or almost every lesson), much as witnessed within the mathematics model.

At G8 some evidence on reform-based practice is reversed, with explaining rather than memorising offering significant association with science achievement scores. The empirical data appears to support the call for explaining, discussion and argumentation as put forward in the literature discussed in Chapter 2. Memorising science facts and principles is non-significant in the G8 model that places a greater emphasis on explaining as a way of interpreting variance in science achievement scores.

Only one of the response categories within the learning environment cluster was significantly associated with achievement, where working on own 'every or almost every lesson' had a reported effect size of $\Delta = 0.17$. The other contributing experiences in this cluster were non-significant, so the data does not support working in groups (as with G4) or lecture-style presentation as features that provide significant explanation of variance in science achievement scores. Although the magnitude of effect sizes in the G8 model are generally smaller, the highest effect of $\Delta = 0.23$ was attributed to the non-significant entry on lecture-style presentation (every or almost every lesson); this suggests there is an association with achievement scores but insufficient support on the current model to make any significant claims.

The pattern and strength of association with achievement for practical and experimental work reported above in G4 model is not repeated with the G8 data. Indeed all of those activities, of watching teacher demonstrate an experiment, conducting an experiment or investigation themselves, and planning an experiment or investigation are non-significant in the G8 model of science achievement. Within the multivariate cluster analysis of those three activities (conduct, watch, and plan), the strongest association with science achievement scores was with conducting experiments, where the effects were all

positive ($\Delta = 0.24$ to $\Delta = 0.41$) with the highest effect linked to conducting experiments every or almost every lesson; this reflects a similar outcome to that reported in G4 where doing experiments was valued in terms of association with achievement outcomes. However in the final model that combined all predictor and control variables in G8 science, none of the responses on practical experiments and investigations provide a significant contribution to the variance in the model. This raises interesting questions over practice and classroom experiences in a sector where the students have the benefit of specialist teachers. The exploratory data analysis shows that seventy five per cent of the students are conducting experiments, seventy two per cent are watching their teacher demonstrate an experiment or investigation, and over sixty per cent are planning experiments or investigations in about half their lessons or more frequently. So although the students are engaging in practical work and experiments there is no visible gain from that effort when those activities are considered alongside other classroom experiences and organisational features of learning and teaching in science.

In the final CPV model of G8 science achievement the strongest positive associations are attributed to opportunities to explain science studied ($\Delta = 0.20$), and to work on problems on their own ($\Delta = 0.17$). This finding would appear to be supportive of reform-based practice, where opportunities to explain science studied is very much at the heart of the reform movement with an emphasis on conceptual understanding and social constructivist approaches to learning as discussed in Chapter 2.

The role of ICT in learning and teaching does not appear to have a positive association with achievement in Science classrooms. There is however one highly significant but small positive effect in G8 ($\Delta = 0.14$) when computer use is witnessed in support of science school work (a few times a year); noting this particular use of computer technology is outwith the classroom setting. There is a significant negative association with science achievement when computers are used more regularly in support of G8 school work (at least once a week) where the effect size is $\Delta = -0.16$. A strong negative effect is noted for computer use in G8 classes, with $\Delta = -0.17$ (about half the lessons) and $\Delta = -0.33$ (for every or almost every lesson).

One other variable from the G8 analysis is highlighted as being significantly associated with science achievement scores. This is within the assessment and feedback cluster along the same lines as reported in mathematics, but in the science model it is review of homework that provides the significant association; quiz or test is non-significant. The reported effects for review of homework are $\Delta = -0.20$ (about half the lessons) and $\Delta = -0.25$ (every or almost every lesson). As above, this indicates the association with

achievement scores is not in line with what might be expected given the conceptual framework and policy initiatives on AifL, assessment and feedback.

In summarising the substantive issues associated with achievement in science, and to address the second part of research question 1 (b), I will cross-reference to the conceptual framework in the same way as handled for Mathematics. Table 7.2-4 presents the principal associations with educational performance in science as reported in $TIMSS_{2007}$, with entries ranked by highest individual effect.

Table 7.2-4: Summary of Associations with Achievement in Science (conceptual link and highest individual effect)

	Positive association	Negative association
Grade 4	Memorise science facts (RPD, Δ=0.36) Do science experiments (ALPA, Δ=0.23) Work on own (LE, Δ=0.22) Watch science experiments (ALPA, Δ=0.21)	Use computers in support of school work (T, Δ= -0.30)
Grade 8	Explain science studied (RPD, Δ=0.20) Work on own (LE, Δ=0.17) Use computer in support of school work a few times a year (T, Δ= 0.14)	Use a computer in science lesson (T, Δ= -0.33) Review our homework (ALPA, Δ= -0.25) Use computer in support of school work at least once a week (T, Δ= -0.16)

Differences across stage within science

Addressing the second half of research question 1 (d), some differences within science are noted across the primary and secondary stages. First concerns the place of practical experiments and investigations in school science. A strong positive association with achievement is noted in G4 for opportunities to do science experiments; there is also a positive association with watching science experiments (a few times a year). In contrast, the equivalent experiences in G8, to conduct experiments or watch teacher demonstrations of science experiments or investigations, are all non-significant in the model of achievement. The data does not support active learning and practical experiments as features that provide explanation of variance in G8 science achievement scores.

Second, in the learning environment cluster, we can see that opportunities for 'working on own' are strongly associated with both stages but more widely impacting on achievement in G4 where two response options report a high effect size; much as reported for mathematics there is only a significant association in G8 when witnessed in every or almost every lesson. The interesting finding in this cluster is that the data does not support

group work as a feature of learning that can explain variance in science achievement at either stage.

Finally there is a contrast in reform-based practice and the place of memorising over explaining. The dominant feature and significant association with achievement in G4 science rests on memorising science facts, with a non-significant association for explaining. Whereas in G8 the associations are reversed, with a positive association between explaining science studied (every or almost every lesson) and achievement, and a non-significant association with memorising science facts and procedures. It would be worth investigating whether teacher knowledge and confidence with science has a bearing on this division where we have specialist teachers in the secondary stages and normally non-specialists teaching science in the primary sector. The latter group may lean towards a traditional memorisation route in preference to engaging in discussion and active learning that would make higher demands on teachers' subject content knowledge. Analysis of association between students' achievement scores and teachers' background knowledge, subject competence, highest qualifications and professional development can be explored; initial analyses along those lines suggest this would potentially address that particular question and is identified for future study.

7.2.3. Similarities and differences across disciplines within stage

Discussion of differences within stages that have not already been reported will address research question 1 (c). The principal associations at G4 that overlap disciplines concern the learning environment (work on own; work in group) and reform-based practice (memorise; explain; active learning and practical activities). Those variables have already been discussed, highlighting a lack of support for any association with achievement when working in a group or explaining in science education, whereas both of those practices are positively associated with achievement in mathematics. Reform-based practices of collaborative work, discussion and explanation of understandings appear to be established as positive experiences for students in mathematics, where significant benefits are reported as a result of exposure to those practices. The same cannot be said for science teaching where the emphasis resides with memorising as discussed in 7.2.2.

Two areas that stand out in G8 are related to assessment and feedback (review homework; have a quiz or test) and reform-based practice (explain, memorise). First it is worth noting that the assessment and feedback practices are negatively associated with achievement for both disciplines. In mathematics education it is the use of quiz or test that is

dominant explanatory variable, with review of homework reported as having a non-significant association with achievement. In Science the opposite pattern is evident with a significant negative association between review of homework and science achievement, and no significant association between quiz or test and achievement. The literature on feedback, formative assessment and AifL all point towards principles that ensure assessment supports learning. This is clearly not happening if achievement is not positively associated with at least some engagement with those practices; the findings indicate students who never experience those practices have higher achievement scores. Survey data such as this can shed light on the effectiveness of policy implementation where there are obvious guidelines and expectations in terms of practice. On reform-based practice it is noticeable that memorising provides explanation of variance in mathematics education, with no significant contribution coming through students explaining their work. The opposite is supported in the science data, potentially reflecting an effect of having a specialist teacher in science but also highlighting that mathematics teachers' practice is not associated with enhanced achievement in the $TIMSS_{2007}$ data. This raises the need for further study of practice, with a focus on the implementation of policies pertaining to discussion, talk and explanation in secondary mathematics settings. Another policy initiative that falls into this category of needing to be more closely evaluated, concerns relating learning to real life, or contextualising learning. Neither discipline provides data that would support an association between contexts or links to daily life and achievement scores.

7.3. Alternative Models

In seeking to address research question 2(a) an analytical framework was developed to support secondary analysis of survey data relevant to practice and policy development in mathematics and science. Consideration was given to a range of models that could be used to analyse the data. There are three aspects of the models that are used to evaluate the merits of each. First, the accuracy of the parameters and estimates of coefficients for each response category; second, the strength of support for any predictor variable to explain variance in the model, as measured by the associated effect size; and third to identify which level of the hierarchy provides explanation of variance by examining the variance partition coefficients and proportion of variance explained, to indicate where efforts might be channelled to improve future practices.

Thinking first about parameters in the model an exploratory data analysis was carried out to determine the scope of the data set and any evident patterns within. On the basis of literature in the field of social and educational research and the data available in the

survey, appropriate control variables were decided upon and used to control the data, as discussed in 7.1, reducing variance in the models of achievement scores so that background noise and identifiable characteristics could be taken into account before proceeding with analyses of substantive issues. The variables outlined in Chapter 4 form the framework for the control variables; these include biographical data (gender, home environment, resources and support for students) and responses on out-of-school interests that can reduce the confounding effect of variations in these variables. Theories of learning and teaching, and policy documentation guided selection of substantive, malleable features of learning and teaching that could provide evidence for policy and practice. For this particular study those variables were restricted to the classroom experiences described and justified in Chapter 4, but they could equally be extended to include teacher data (biographical data, highest qualification, subject knowledge /confidence, engagement in professional development etc.) or school data (management structures, leadership, organisation of classes, etc.) that are all contained within the $TIMSS_{2007}$ survey data set. Once predictor variables were confirmed as potential contributors through inspection of diagnostic analyses they were modelled against achievement scores. Educational practices and pedagogies particular to mathematics and science were considered separately as discussed in 7.2.

Given the nature of survey data collection through matrix sampling design and the fact that achievement scores are imputed as plausible values, particular care had to be taken over methods used to generate parameter estimates and associated standard errors. First the structure of the data collection sampling technique dictated using multilevel modelling techniques to accommodate the hierarchical structure of the nested data. Second the imputation process and use of plausible values called on the evaluation of alternative models to determine the most appropriate approach. The initial plan was to evaluate the merits of extending the multilevel structure to model plausible value as unit of analysis and hence to factor in and take account of the variance within and between the plausible values; this is the multilevel plausible value (MLPV) method reported above. The hypothesis was this approach could potentially by-pass the need to pursue Rubin's Rules for combining multiply imputed estimates. My argument was that the standard errors computed through MLPV method would take account of the variance attributed to the imputation process by nature of the fact that PVs were factored into the model. However, the analyses of different models in Chapter 6 alerted me to the fact that although the MLPV method partitioned variance at plausible value level, identifying the proportion of variance in the model attributed to the imputed data, the calculation of parameter estimates and standard errors did not take account of the imputation process or variance at that level; it treated the plausible values in the

multilevel model as *bona fide* data and ignored the fact that these data had an error component from the imputation process. The four models of achievement scores under consideration took different dependent variables for the achievement score: first the MLPV model that took PV as the unit of analysis in a long data file that included replicate data for students to align with each PV score; second the average of imputed achievement scores modelled as the dependent variable (AVPV); third a single PV as the dependent variable (SPV); and finally comparison of all of the above with Rubin's recommended method of analysis that combines the findings from the five separate models computed with each PV taken as the dependent variable (CPV).

The empirical analyses using these alternative models address research question 2(b). Much as expected, the point estimates using MLPV model are broadly in line with those computed through CPV model that uses Rubin's rules for multiple imputation; the method of calculation of point estimates for each parameter is based on the average of plausible values. However, the standard errors within the MLPV and AVPV models are underestimated and therefore smaller than those computed using CPV model that specifically takes account of within imputation and between imputation variances. This leads to the conclusion that CPV model is the only approach that generates standard errors that take adequate account of the imputation process; SPV model has inflated standard errors that reflect within imputation variance, but such a model cannot factor in any between imputation variance since it only considers a single PV, therefore SPV model also underestimates the standard errors for parameter estimates. A consequence of this is that the CPV model weakens support for conclusions on which predictors explain variance in the models of achievement. There is minimal change on effect sizes, noting their computation is not dependent on the standard error.

In determining which method of analysis is most appropriate in any given circumstance I conclude that it should be guided by the intended purpose of the analysis. From the comparisons summarised in Chapter 6, highly significant associations may reasonably be identified using either a 3-level multilevel model that takes the average of the plausible values as the dependent variable (AVPV model) or a 4-level structure with plausible value as unit of analysis (MLPV model). Substantive conclusions can be drawn on which aspects of the data provide the greatest explanation of variance as measured by effect size, but models based on AVPV or MLPV will definitely under-estimate standard errors and as such any marginally significant findings should be treated with caution. Additionally, the AVPV approach fails to offer a true breakdown of variance partitions at school, teacher and student levels. If that was of interest in addition to identifying principal associations with

achievement, then the MLPV model would provide a fuller picture. The four-level structure identifies proportion of variance explained at student, teacher and school levels of the model, setting aside the variance from imputed plausible values. For instance there were quite different patterns of explanation across sectors. In primary education a third of the variance in the model of achievement was explained at student-level; around half of the teacher-level variance was accounted for; and higher proportions of school-level variance was explained, with nearly eighty per cent of G4 mathematics and two thirds of G4 science variance accounted for in the full models for this stage. The bulk of unexplained variance is left at student-level of the model with small proportions of total variance attributed to school or teacher levels as detailed in Appendix 6.3.

At G8, in the secondary sector, the organisational structures of the school system can impose a different pattern of distribution of variance across the levels of the models. For instance in G8 mathematics only a quarter of unexplained variance is attributed to student level within classes, compared to nearly sixty per cent of the total variance that is aligned with teachers; this will reflect to some extent on setting within classes, reducing within class/between student variance and increasing within school/between teacher variance in the model. In G8 science a high proportion of variance is attributed to school-level of the model, with almost one fifth of total variance determined to be between schools. This is potentially concerning in that it indicates high variation between schools and therefore an inequitable experience for students dependent on school attended. Within those schools the bulk of variance is at student level within classes, which may imply there is less setting by ability in science classes compared to mathematics; this is conjecture because that information is not recorded in the $TIMSS_{2007}$ data set. The MLPV data can tell which level of the structure explains variance that is attributed to particular practices and experiences, and where unexplained variance remains within the model for future analyses. In secondary school science the modelled predictor and control variables explain nearly seventy per cent of the between school variance, over half of the teacher-level variance, and just under a third of the student variance; the latter proportions being very comparable to those witnessed in G4. In secondary mathematics the model provided less explanation than above, with forty three per cent of school level variance accounted for, thirty per cent of teacher variance, and less than a fifth of student-level variance being explained through those predictor and control variables. A high level of variance relative to the other models remains unexplained in the G8 mathematics model.

The intentions behind my study of $TIMSS_{2007}$ data are two fold, first to identify principal associations with educational performance of students, and second to advocate

methods of analysis that could be widely used in order to maximise the use and application of survey data that use plausible value methodology. Using MLPV or AVPV would satisfy both of those intentions, with the rider that marginally significant findings should be treated with caution in the knowledge that standard errors are definitely under estimated when computed using those models; to strengthen arguments a more robust calculation of standard errors would be required, calling for a recalculation of parameter estimates using CPV model (Rubin's rules) i.e. if particular substantive issues are of interest to the researcher then findings need to be accurately confirmed through robust standard errors. Analysis of effect sizes being used as the guiding arbiter in developing a suitable model of achievement scores will be part of a further investigation of these models, rather than relying on significance levels and p-values as the primary measure of impact.

7.4. Conclusions and Recommendations

The proposed research was quite ambitious in that the data sets were so rich that I have only really scratched the surface of what could be explored. That said the substantive findings are of interest to the profession and should be followed through with practitioners to jointly reflect on what measures and future actions are appropriate. My view is that this type of correlation analysis raises questions for the policy and practice communities to reflect on. There are no quick fix solutions, and indeed some of the issues raised are not going to be directly applicable to all, which is why dialogue is required and the profession needs to fully consider evidence from sample surveys like TIMSS, PISA and SSLN as they pursue self-evaluation and plan ahead for their learners. The analysis has highlighted some long-standing issues in STEM education such as gender differences and the impact of home environment on learner's potential achievements as significant explanatory factors, but also identifies recent policy issues that would benefit from discussion within the profession, including features of reform-based practice and discussion, active learning and practical activities, use of technology, and the practice of assessment and feedback in learning and teaching. Reflection on the findings in this study can pave the way for discussion of pedagogical practices, their association with achievement in mathematics and science education, and how they relate to policy developments. The evidence gathered on methods of analysis are also worthy of further dissemination and development as research, policy and practice communities consider deeper analyses of sample surveys based on imputed data.

7.4.1. Implications for policy and practice

This thesis has provided a much-needed contribution to the literature on reform practices and how they are associated with achievement. Prior to this study much of what has been published on survey data has been of a high-level summary format, providing international comparisons, league tables of country performances, and analyses of trends over time, but without identifying associations within the data or measures of impact that could be linked to particular practices. For instance the typical reporting format states findings in prose, painting a picture of the evidence collected through the survey but not necessarily informing the audience beyond the facts:

> Students in the United Kingdom score 494 points in mathematics, on average – at the OECD average and comparable with the Czech Republic, Denmark, France, Iceland, Republic of Ireland, Latvia, Luxembourg, New Zealand, Norway and Portugal. Mean performance in mathematics has remained unchanged since 2006 and 2009. (OECD, 2013)

> In terms of assessment, two thirds (66%) of S2 pupils have teachers who place a major emphasis on classroom tests in maths, no different to the international average, whilst just under half (47%) of pupils have teachers who place a major emphasis on the teacher's own professional judgement, again no different to the international average. 17% of pupils have teachers who place a major emphasis on national or regional tests, below the international average of 27%. (Horne *et al.*, 2008)

What does 494 points mean in PISA and what relationship if any is there between teachers' emphases on particular practices and student achievement scores in TIMSS? Clearly there are opportunities to undertake deeper analyses of policy and practice in relation to achievement scores reported through surveys and this thesis has sought to make inroads on such analyses. The researcher can identify aspects of particular educational interest and use the evidence from sample survey data to inform practitioner and policy communities of outcomes and associations with achievement. A broad range of outcomes and experiences can be considered with scope to reflect on policy-as-implemented as well as evaluating associations between pedagogical practices and achievement scores. This can include commentary on any gaps between policy aims and practitioners' implementation, highlighting the importance of non-significant predictors in models of achievement as well as focusing on those predictors that offer significant explanation of variance and strong effects in the models. For instance this exercise provides an opportunity to reflect on the policy context and implementation of principles and practices promoted by *Curriculum for Excellence* in Scotland, and to evaluate the efficacy of practices as promoted within research on mathematics and science learning and teaching. The findings lead me to conclude that

implementation of those policies is far from complete and indeed the associations between practice and achievement are not always in line with research and policy claims. The details of those shortfalls and deviations from research are summarised in Chapter 6; the aspects of learning, teaching and assessment that practitioners and other stakeholders would do well to reflect on, giving due consideration to the associated research to guide development of practices that enhance student achievement.

Several issues reported in this study prompt wider and further research, either through additional secondary analyses of TIMSS survey data that focus on teacher and school levels of the data, or through primary data collection on specific issues that would benefit from focused experimental or observational studies and in-depth qualitative analyses to expand on what can be gleaned from the survey responses. I would highlight four such issues that are worthy of exploring. First, additional secondary analyses of achievement data with school- and teacher-level variables could provide evidence of association between teacher's qualifications, self-confidence in subject knowledge, and opportunities for professional development with science achievement scores. Findings suggest this to be of particular relevance in science education, noting a lack of opportunity for explaining science studied in G4 as a feature that is potentially linked to teacher confidence as discussed in Section 7.2.2 and reported in Appendix 4.3. Secondary analysis of teacher- and school-level variables provides a rich source of information to explain variances in student achievement, picking up on teacher quality as above and accounting for school characteristics including factors such as school composition, leadership, parental engagement, and resources. Second, the use of technologies in learning and teaching and how that culture change identified in the Scottish Government's objectives for ICT can really be taken forward. An understanding of the negative association reported between ICT and achievement in mathematics and science is required before practices can be reviewed and developed in line with policy aspirations. An experimental study on the use of ICT and its impact on learning and achievement would be feasible, with the option for a control group given the staged availability and implementation of particular platforms and supporting educational software. Third there remains a question over the role, focus and purpose of science experiments and investigations. The literature has tabled a range of views on the benefits of experiments and scientific method but it was interesting to note a shift in emphasis and strength of association with practical and active learning evidenced in the primary environment (G4) compared to the more traditional focus pursued in the secondary stages (G8). Follow-up analyses should seek to ascertain what lies behind those findings to identify practices that would be beneficial for students. If practical experiments and investigations are deemed to be a key

element of effective practice then further studies are needed to establish how practitioners can support learners to maximise the impact and effect of planning, watching or conducting science experiments and investigations. Fourth, there is a need for further study of the cluster of variables on assessment and feedback to establish why existing practices are not enhancing learning and achievement as claimed in the literature. This will require an analysis of *purpose* as well as *practice* to clarify what practitioners are seeking to gain from those activities in order to make sense of the negative association between feedback and achievement as reported through opportunities to review homework, and to complete tests and quizzes in G4 and G8 study of mathematics and science. A major limitation of questionnaire survey methods is that they cannot fully reflect the complexity of the classroom and teachers' practice in that there will be a range of ways in which common practices can be implemented – the details over how basic practices are deployed or implemented will determine level of impact on learners e.g. lecture-style presentations could be passive or interactive; review of homework may make low or high demands on students' conceptual understanding; memorisation of facts and procedures could be purely rote and instrumental or fully integrated, connected and relational. Where there are unexplained variances or conflicting messages in the findings reported in this thesis it would be beneficial to investigate the issues through alternative measures and different research instruments.

On the methods of analysis, I believe my exploration into scaling combined responses is worthy of wider application where response scales have an underlying continuous measure. Transforming the scales in line with Snell's recommendation makes fuller use of the data and has the potential for a more refined analysis with derived variables considered as continuous variables in models of achievement scores. This will be further investigated with a range of different response scales, drawing parallels with the work of Page-Bucci (2003) who considers a continuous metric through use of a computer slider during data collection, and evaluating the merits of transforming ordinal polytomous response categories into a continuous metric for subsequent analysis.

The empirical data has also provided convincing evidence of practice and insight to the interpretation of analyses based on plausible value methodology. The multilevel models suitably address the hierarchical structure of the data and as such should be used by all analysts for this type of clustered data, where observations within a cluster cannot be assumed to be independent. Any disregard for the sample design will underestimate the true sampling variability as set out in Chapter 3. The compounding consequence of handling imputed data opened an opportunity to investigate the merits of a multilevel plausible value model (MLPV) that could be compared with findings through appropriate alternatives,

including the currently recommended route that uses Rubin's rules to combine parameter estimates. The long data format with plausible value as unit of analysis had benefits in that the partitioning of variance was clearly maintained, as opposed to having the quite sizeable variance between plausible values being absorbed into one of the higher levels of the structure. However, the comparative findings highlighted deficiencies in my proposed method. The MLPV model underestimated the standard errors for parameter estimates; facing similar criticisms as levelled at analysts who use the average of plausible values as the dependent variable (AVPV). Although I concluded that using the MLPV or AVPV model can provide the main messages from the data, any marginal issues will not be accurate. This leaves the combined plausible value model (CPV) as the only valid route that will correctly estimate standard errors to accompany point estimates computed in the models of achievement data. The important message here is that methods must fit the data. In this case and in other large scale educational survey data collections this means accommodating the sampling framework and method of capturing cognitive data, i.e. handling clustered data and plausible values based on multiple imputation processes.

7.4.2. Recommendations

The title of this thesis refers to 'using and applying' survey data and to this end the recommendations are directed towards three stakeholder groupings who have influence over learners and their classroom experiences, namely: educational researchers, policy makers, and classroom practitioners. Each group has a role to play in using and applying survey data to inform policy and practice, and to ultimately impact of students' achievement:

Educational researchers

1. Analyse and disseminate findings that go beyond descriptive accounts of findings
    a. present correlational analyses that can begin to quantify effects and associations with achievement scores
    b. develop coherent and meaningful analytical frameworks that combine background characteristics, contextual variables and explanatory variables guided by policy and literature; building on framework within thesis
    c. argue case for inclusion of data on students' prior learning to improve analytical interpretation of teacher and school influences on achievement; and to permit value-added analyses as appropriate
    d. dissemination strategy to reach relevant stakeholders
2. Undertake in-depth analyses of survey data using appropriate methods that take account of sample structure, design and method of data collection
    a. multilevel modelling technique for clustered data
    b. combined plausible value method using Rubin's rules to handle imputed data from matrix sampling design
    c. report effect sizes as well as significance levels to quantify impact of practice/ experience
3. Plan and implement additional research to complement survey analyses
    a. focused experimental studies on the role of ICT in learning and teaching to ascertain impact if any on achievement; a particular focus on association with achievement as a separate outcome from any findings on the affective domain
    b. secondary analyses of teacher- and school- level variables to explain variances in achievement that can be attributed to teacher quality, school context and ethos; exploiting the MLM structure for analysis

c. observational analyses to follow up on initial findings in respect of science practical activities, assessment and feedback; establish what lies behind the non-significant and negative associations reported through survey analyses

Policy makers

4. Use survey data as a contributing component in evaluation of policy initiatives
    a. interrogate direction of association with achievement
    b. take opportunity to justify and exemplify policy-in-practice that does not show significant association with achievement as expected or as claimed in research literature; non-significance is equally important for dissemination of practice if argued to have a positive association with achievement scores
5. Continue to commission appropriate survey data that can service needs in providing international benchmarking, a national picture of practice, and robust national monitoring through sample surveys that reduce demands on students
    a. SSLN for primary and secondary stages (P4, P7, S2)
    b. TIMSS for primary and secondary stages (P5, S2)
    c. PISA for secondary only (S3 ~age 15years)
6. Disseminate digestible and useable outputs for educational researchers and practitioner audiences, focusing on how the investment in survey data collection can be used to inform and refine policy advice; this will require analyses that go beyond descriptive accounts of the data to maximise use and application of survey data to impact on students' achievement

Classroom practitioners

7. To engage with findings from survey data; engagement should take the form of collaborative discussion and reflection at a local level to reflect on frequency of opportunity afforded to students and on how specific practice is pursued by self and colleagues.
8. Reflective practice that is built upon research informed practice – literature and survey data combined to inform and contribute to professional learning as set out in *Professional Update* for registered teachers in Scotland:

> Professional learning is what teachers do to ensure their professional knowledge and practice is informed, up-to-date and stimulating. It is important that professional learning provides

> rich opportunities for teachers to develop and enhance their professional knowledge and practice, in order to progress the quality of learning and teaching and school improvement.
> (General Teaching Council for Scotland (GTCS), 2014)

Each of those recommendations seeks to raise the profile of survey data by building on findings in the thesis to influence the way reform-based practice, discussion, active learning, practical activities, technology, assessment and feedback, real life contexts and students' learning environment are taken forward to enhance achievement. This will entail re-focusing attention on core elements of pedagogy and encouraging practitioners to reflect on how they go about implementing recent policy initiatives.

In recommending continued investment in large scale surveys such as TIMSS, PISA and SSLN, I acknowledge that between starting and completing this study the Scottish Government withdrew from TIMSS and decided only to continue participation with the PISA (and PIRLS) studies. That makes $TIMSS_{2007}$ the final wave of the survey as far as Scotland is concerned. I'd argue this to be a major loss in the overall purposes set out for participation in international monitoring of performance and standards, because there will be no benchmarking facility for primary education or opportunity to capitalise on STEM analyses such as those pursued in this thesis. Continuation with PISA provides an international perspective for Scottish education, but only at age 15, which is towards the end of the compulsory phase of education in Scotland. This cannot therefore provide any evidence of practice or insights to implementation of policy within the core period of education that spans the first ten years of formal education. The argument from the Scottish Government that one international benchmarking exercise is sufficient is disappointing, especially since PISA is targeted at those leaving the education system. It would be considerably richer to have the opportunity to explore robust internationally gathered data on learners' experiences and achievements whilst still in their primary and early secondary stages of school.

7.5. Future actions

My intention is to continue working on aspects of TIMSS data that have yet to be uncovered and to seek collaboration with colleagues from the Netherlands where there are interesting STEM initiatives that could be linked to the survey variables and analyses of students' achievement data. One area I intend to explore further concerns the data on teacher development, given my professional role as a teacher educator and wider involvement in continuing professional development; my initial exploration into this branch of $TIMSS_{2007}$ was promising but that could not be fitted into this thesis. The international element of the data was also useful in shedding light on the different background characteristics that came through EDA on a cross-national platform. However, an in-depth knowledge of the different cultures and local communities is essential for the analyst to interpret the data – if nothing else to get accurate data on what 'country-specific' questions were included and why; hence the desire to work in collaboration with colleagues in host country. The analytical framework developed in this thesis can be transferred to other national contexts, permitting analysis of the next wave of TIMSS to confirm substantive claims identified in this thesis and to consolidate methodological developments, albeit with analyses based on countries other than Scotland. The methodological developments discussed in this thesis can also be deployed in analyses of PISA data. The structure and management of PISA data follows similar lines to those witnessed in TIMSS, with clustered samples and imputed data presented through five plausible values for each student in the study. The Scottish context and measures of achievement can therefore be researched and followed up through an analysis of $PISA_{2012}$ using the same methodology as reported here. However, since there are no overlapping stages with the national survey of literacy and numeracy (SSLN) the analysis of PISA data will not carry the same degree of linkage to the SSLN that would have been afforded through a parallel analysis of $TIMSS_{2011}$; hence the disappointment that Scotland is no longer participating in the IEA survey that is at the heart of this thesis.

8. References

ABERDEIN, A. 2005. The Uses of Argument in Mathematics. *Argumentation,* 19**,** 287-301.

ABRAHAMS, I. 2011. *Practical Work in Secondary Science: A minds-on Approach,* London, Continuum.

ABRAHAMS, I. & REISS, M. J. 2012. Practical work: Its effectiveness in primary and secondary schools in England. *Journal of Research in Science Teaching,* 49**,** 1035-1055.

AIRASIAN, P. W. & WALSH, M. E. 1997. Constructivist cautions. *Phi Delta Kappan,* 78**,** 444.

ALEXANDER, R. 2006. *Towards Dialogic Teaching: rethinking classroom talk,* Cambridge, Dialogos.

ALSTON, C., KUHNERT, P., LOW, S., MCVINISH, R. & MENGERSEN, K. 2005. Bayesian Model Comparison: Review and Discussion. Available: http://iase-web.org/documents/papers/isi55/Alston-Kuhnert-Low_Choy-McVinish-Mengersen.pdf.

ANDREWS, R. 2009. *The importance of argument in education,* London, Institute of Education.

ARC CENTER 2003. *The ARC Center Tri-State Student Achievement Study,* Chicago, National Science Foundation.

ATHERTON, J. S. 2011. *Learning and Teaching; Constructivism in Learning* [Online]. Available: www.learningandteaching.info/learning/constructivism.htm [Accessed 30 November 2011.

ATWOOD, S., TURNBULL, W. & CARPENDALE, J. I. M. 2010. The Construction of Knowledge in Classroom Talk. *Journal of the Learning Sciences,* 19**,** 358-402.

AUSUBEL, D. P. 1968. *Educational Psychology: A Cognitive View,* New York, Holt, Rinehart & Winston.

BARNES, D., THOMSON, J. & WATSON, K. 1978. *Language in the Classroom,* Melbourne, Applied Linguistics Association of Australia.

BEATON, A. E. 1987. Implementing the New Design: The NAEP 1983-84 Technical Report. National Assessment of Educational Progress, Princeton, NJ.

BEATON, A. E. & ZWICK, R. 1992. Overview of the National Assessment of Educational Progress. *Journal of Educational Statistics,* 17**,** 95-109.

BEREITER, C. 1994. Constructivism, Socioculturalism, and Popper's World 3. *Educational Researcher,* 23**,** 21-23.

BIDWELL, C. E. & KASARDA, J. D. 1980. Conceptualizing and Measuring the Effects of School and Schooling. *American Journal of Education,* 88**,** 401-430.

BIEMANS, H. A. & SIMONS, P. R.-J. 1995. How to use preconceptions? The contact strategy dismantled. *European Journal of Psychology of Education,* 10**,** 243-259.

BIESTA, G. 2007. WHY "WHAT WORKS" WON'T WORK: EVIDENCE-BASED PRACTICE AND THE DEMOCRATIC DEFICIT IN EDUCATIONAL RESEARCH. *Educational Theory,* 57**,** 1-22.

BINDER, D. A. & ROBERTS, G. R. 2003. Design-based and Model-based Methods for Estimating Model Parameters. *In:* CHAMBERS, R. L. & SKINNER, C. J. (eds.) *Analysis of Survey data.* Chichister: Wiley.

BLACK, P. J., WILIAM, D. & KING'S COLLEGE (LONDON ENGLAND). DEPT. OF EDUCATION AND PROFESSIONAL STUDIES. 1998. *Inside the black box : raising standards through classroom assessment,* London, Dept. of Education & Professional Studies, King's College London.

BRIDGES, D., SMEYERS, P. & SMITH, R. 2009. *Journal of philosophy of education.Evidence-based education policy: what evidence? what basis? whose policy?,* Chichester, West Sussex, Wiley-Blackwell.
BROUSSEAU, G. 1997. *Theory of didactical situations in mathematics,* Dordrecht, Kluwer.
BROWN, A. S. & BROWN, L. L. 2007. What are Science & Math Test Scores Really Telling U.S.? *The Bent of Tau Beta Pi* [Online], Winter 2007. Available: http://www.tbp.org/pubs/Features/W07Brown.pdf.
BROWN, J. S., COLLINS, A. & DUGUID, P. 1989. Situated Cognition and the Culture of Learning. *Educational Researcher,* 18**,** 32-42.
BROWNE, W. 2012. MCMC estimation in MLwiN.
BULL, A., GILBERT, J., BARWICK, H., HIPKINS, R. & BAKER, R. 2010. Inspired by science: A paper commissioned by the Royal Society and the Prime Minister's Chief Science Advisor. Wellington: New Zealand Council for Educational Research.
BUS, A. G., VAN IJZENDOORN, M. H. & PELLEGRINI, A. D. 1995. Joint Book Reading Makes for Success in Learning to Read: A Meta-Analysis on Intergenerational Transmission of Literacy. *Review of Educational Research,* 65**,** 1-21.
BUXTON, L. 1978. Four levels of understanding. *Mathematics in Schools,* 7**,** 36.
BYERS, V. & HERSCOVICS, N. 1977. Understanding School Mathematics. *Mathematics Teaching,* 81**,** 24-27.
BYRNE, D. & SMYTH, E. 2010. Behind the Scenes? A study of parental involevment in post-primary education. Dublin: Liffey Press & ESRC.
CADIMA, J. & JOLLIFFE, I. T. 1995. Loading and correlations in the interpretation of principle compenents. *Journal of Applied Statistics,* 22**,** 203-214.
CAHN, S. M. 2009. *Philosophy of education : the essential texts,* New York, Routledge.
CERINI, B., MURRAY, I. & REISS, M. J. 2003. Student Review of The Science Curriculum. London: Institute of Education.
CHAPMAN, M., WYKES, C. & COLLEGE, C. S. 1996. *Plain figures*, Stationery Office.
CLARKSON, S. G. & WRIGHT, D. K. 1992. An appraisal of practical work in science education. *School Science Review,* 74**,** 39-42.
COCKCROFT, W. H. & COMMITTEE OF INQUIRY INTO THE TEACHING OF MATHEMATICS IN SCHOOLS 1982. Mathematics counts :report of the Committee of Inquiry into the Teaching of Mathematics in Schools under the chairmanship of W.H. Cockcroft.
COE, R. 2002. It's the Effect Size,Stupid: Waht effect size is and why it is important. *British Educational Research Association.* University of Exeter: Education-Line.
COHEN, J. 1969. *Statistical Power Analysis for the Behavioural Sciences,* New York, Academy Press.
CONDIE, R., LIVINGSTON, K. & SEAGRAVES, L. 2005. The Assessment is for Learning Programme: An Evaluation Edinburgh: Scottish Executive Education Department.
CONFREY, J. 1990. A Review of the Research on Student Conceptions in Mathematics, Science, and Programming. *Review of Research in Education,* 16**,** 3-56.
CONFREY, J. & KAZAK, S. 2006. A thirty-year reflection on constructivism in mathematics education in PME. *In:* GUTIERREZ, A. & BOERO, P. (eds.) *Handbook of Research on the Psychology of Mathematics Education.* Sense Publishers.
CORREA-CHAVEZ, M. & ROGOFF, B. 2005. Cultural Research has transformed our ideas of cognitive developmenty. *International Journal of Behavioral Development,* 29.
COX, E. P. 1980. The Optimal Number of Response Alternatives for a Scale: A Review. *Journal of Marketing Research,* 17**,** 407-422.
CUBAN, L. 2001. *Oversold and underused: computers in the classroom,* Massachusetts, Harvard University Press.

DEARY, I. J., THORPE, G., WILSON, V., STARR, J. M. & WHALLEY, L. J. 2003. Population sex differences in IQ at age 11: the Scottish mental survey 1932. *Intelligence,* 31**,** 533-542.

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39**,** 1-38.

DEWEY, J. 1916. *Democracy and Education: An Introduction to the Philospohy of Education,* New York, Free press.

DITTRICH, R., FRANCIS, B., HATZINGER, R. & KATZENBEISSER, W. 2007. A paired comparison approach for the analysis of sets of Likert-scale responses. *Statistical Modelling,* 7**,** 3-28.

DRIVER, R. 1983. *The Pupil as Scientist?,* Milton Keynes, Open University Press.

DRIVER, R. 1989. Students' conceptions and the learning of science. *International Journal of Science Education,* 11**,** 481-490.

DRIVER, R., ASOKO, H., LEACH, J., MORTIMER, E. & SCOTT, P. 1994. Constructing Scientific Knowledge in the Classroom. *Educational Researcher,* 23**,** 5-12.

DUCKWORTH, E. 1987. *'The having of wonderful ideas' and other essays on teaching and learning,* New York, Columbia University, Teachers College.

ECONOMIC & SOCIAL RESEARCH COUNCIL. 2013. *Secondary Data Analysis Initiative Phase 2* [Online]. Available: http://www.esrc.ac.uk/funding-and-guidance/funding-opportunities/26605/secondary-data-analysis-initiative-phase-2november.aspx.

EDUCATION SCOTLAND 2009a. Curriculum for Excellence.

EDUCATION SCOTLAND. 2009b. *Experiences and Outcomes* [Online]. Available: http://www.educationscotland.gov.uk/learningteachingandassessment/curriculumareas/mathematics/eandos/index.asp.

EDUCATION SCOTLAND. 2010a. *Principles and practice: mathematics* [Online]. Available: http://www.educationscotland.gov.uk/learningteachingandassessment/curriculumareas/mathematics/principlesandpractice/index.asp.

EDUCATION SCOTLAND. 2010b. *Principles and practice: sciences* [Online]. Available: http://www.educationscotland.gov.uk/learningteachingandassessment/curriculumareas/sciences/principlesandpractice/index.asp.

EHRENBERG, A. S. C. 1986. *A primer in data reduction :an introductory statistics textbook,* Chichester, Wiley.

EURYDICE NETWORK 2011. Mathematics Education in Europe: Common Challenges and National Policies Brussels: Education, Audiovisual and Culture Executive Agency

FAIELLA, I. 2010. The use of survey weights in regression analysis. Bank of Italy, Economic Research and International Relations Area.

FENNEMA, E. 2002. Gender Equity for Mathematics and Science. *GEMS.*

FIDLER, F. 2010. The American Psychological Association Publicatuion Manula SixthEdition: Implications for Statistics Education. *In:* READING, C. (ed.) *Data and context in statistics educatoin: Towards an evidecne-based society.* Voorburg, Netherlands: Int. Statistical Institute.

FORD, M. J. 2010. Critique in academic disciplines and active learning of academic content. *Cambridge Journal of Education,* 40**,** 265-280.

FOSNOT, C. T. 1996. Constructivism :theory, perspectives, and practice /Catherine Twomey Fosnot, editor.

FOSNOT, C. T. & PERRY, R. S. 2002. Constructivism: A Psychological Theory of Learning. *In:* FOSNOT, C. T. (ed.) *Constructivism: Theory, perspectives, and practice.* New York: Teachers College Press.

FOY, P., GALIA, J. & LI, I. 2008. Scaling the Data from the TIMSS 2007 Mathematics and Science Assessments *In:* OLSEN, J. F., MARTIN, M. O. & MULLIS, I. V. S. (eds.) *TIMSS 2007 Technical Report.* Chestnut Hill, MA: TIMSS & PIRLS International Study Centre, Boston College.

FOY, P. & OLSEN, J. F. 2009. *TIMSS 2007 User Guide: Supplement 3,* Chestnut Hill, TIMSS & PIRLS International Study Centre.

FULLAN, M. 1993. *Change forces :probing the depths of educational reform /Michael Fullan,* London :, Falmer.

FUSON, K. C., CARROLL, W. M. & DRUECK, J. V. 2000. Achievement Results for Second and Third Graders Using the Standards-Based Curriculum Everyday Mathematics. *Journal for Research in Mathematics Education,* 31**,** 277-295.

GAITO, J. 1980. Measurement Scales and Statistics: Resurgence of an Old Misconception. *Psychological Bulletin,* 87**,** 564-567.

GARDNER, P. L. 1975. Scales and Statistics. *Review of Educational Research,* 45**,** 43-57.

GEELAN, D. R. 1997. Epistemological Anarchy and the Many Forms of Constructivism. *Science & Education,* 6**,** 15-28.

GEIST, E. & KING, M. 2008. Different,Not Better: Gender Differences in Mathematics Learning and Achievement. *Journal of Instructional Psychology,* 35**,** 43-52.

GENERAL TEACHING COUNCIL FOR SCOTLAND (GTCS), 2014. *Professonal Update is coming* http://www.gtcs.org.uk/web/FILES/professional-development/professional-update-leaflet-0414.pdf

GIVON, M. M. & SHAPIRA, Z. 1984. Response to Rating Scales: A Theoretical Model and Its Application to the Number of Categories Problem. *Journal of Marketing Research,* 21**,** 410-419.

GLUCKMAN, P. 2011. *Looking Ahead: Science Education for the Twenty-First Century,* Auckland, Office of the Prime Minister's Sciecne Advisory Committee.

GONZALEZ, J. M. & ELTINGE, J. L. Multiple Matrix Sampling: A Review. American Statistical Association, Proceedings of the Section on Survey Research Methods, 2007.

GORSUCH, R. L. 1983. Factor analysis /Richard L. Gorsuch. Hillsdale, N.J. ;London :: Erlbaum.

GREK, S., LAWN, M., LINGARD, B., OZGA, J., RINNE, R., SEGERHOLM, C. & SIMOLA, H. 2009a. National policy brokering and the construction of the European Education Space in England, Sweden, Finland and Scotland. *Comparative Education,* 45**,** 5-21.

GREK, S., LAWN, M. & OZGA, J. 2009b. Production of OECD's 'Programme for International Student Assessment (PISA)'. *Know+Pol Working Paper 11.* Louvain-la-Neuve: Université Catholique de Louvain.

HALL, B. 2012. Bayesian Inference. Available: http://www.statisticat.com/laplacesdemon.html.

HALLAM, S., KIRTON, A., PFEFFERS, J., ROBERTSON, P. & STOBART, G. 2004. Evaluation of Project 1 of the Assessment is for Learning development programme: Support for professional practice in formative assessment London: Institute of Education, University of London.

HAMBLETON, R. K. & COOK, L. L. 1977. Latent Trait Models and Their Use in the Analysis of Educational Test Data. *Journal of Educational Measurement,* 14**,** 75-96.

HARLEN, W. 2006. *Teaching, Learning and Assessing Science 5-12,* London, Sage.

HARRIS, D. 1989. Comparison of 1-, 2-, and 3-Parameter IRT Models. *Instructional Topics in Educational Measurement,* Spring.

HATTIE, J. 2009. *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*, Routledge.

HATTIE, J. & TIMPERLEY, H. 2007. The Power of Feedback. *Review of Educational Research,* 77**,** 81-112.

HAUSER, R. M. 1994. Measuring Socioeconomic Status in Studies of Child Development. *Child Development,* 65**,** 1541-1545.
HAYLOCK, D. & COCKBURN, A. 2003. *Understanding Mathematics in the Lower Primary Years,* London, Sage.
HAYWARD, L., SIMPSON, M. & SPENCER, E. 2005. Assessment is for Learning: Exploring programme success. Available: http://www.highland.gov.uk/NR/rdonlyres/11368A71-C450-40DC-8FF4-1285D6749B3B/0/B1AifLExploringProgrammeSuccessS6.pdf.
HENDERSON, S. & CUNNINGHAM, E. 2011. Curriculum reform in Scotland: principles and practice. *In:* HUDSON, B. & MEINART, M. (eds.) *Beyond Fragmentation: Didactics, Learning and Teaching in Europe.* Leverkusen-Opladen: Barbara Budrich Publishers.
HER MAJESTY'S INSPECTORATE OF EDUCATION (HMIE) 2001a. Standards and Quality in Primary and Secondary Schools: 1998-2001. Edinburgh: Stationery Office.
HER MAJESTY'S INSPECTORATE OF EDUCATION (HMIE) 2001b. Standards and Quality in Primary Schools: Mathematics 1998-2001. Edinburgh: Stationery Office.
HER MAJESTY'S INSPECTORATE OF EDUCATION (HMIE) 2005a. Improving Achievement in Mathematics in Primary and Secondary Schools. Improving Series. Her Majesty's Inspectorate of Education.
HER MAJESTY'S INSPECTORATE OF EDUCATION (HMIE) 2005b. Improving Achievement in Science in Primary and Secondary Schools. Improving Series. Her Majesty's Inspectorate of Education.
HER MAJESTY'S INSPECTORATE OF EDUCATION (HMIE) 2006. Improving Scottish Education. Livingston: HMIE.
HER MAJESTY'S INSPECTORATE OF EDUCATION (HMIE) 2008. Science: A Portrait of Current Practice in Scottish Schools. Her Majesty's Inspectorate of Education.
HER MAJESTY'S INSPECTORATE OF EDUCATION (HMIE) 2009. Improving Scottish Education. Livingston: HMIE.
HER MAJESTY'S INSPECTORATE OF EDUCATION (HMIE) 2010. *Learning Together: Mathematics*, Her Majesty's Inspectorate of Education.
HIDI, S. & HARACKIEWICZ, J. M. 2000. Motivating the Academically Unmotivated: A Critical Issue for the 21st Century. *Review of Educational Research,* 70**,** 151-179.
HIEBERT, J. & CARPENTER, T. 1992. Learning and teaching with understanding. *In:* GROUWS, D. (ed.) *Handbook of Research in Mathematics Teaching and Learning.* New York: acmillan.
HODSON, D. 1996. Practical work in school science: exploring some directions for change. *International Journal of Science Education,* 18**,** 755-760.
HOFSTEIN, A. & LUNETTA, V. N. 1982. The Role of the Laboratory in Science Teaching: Neglected Aspects of Research. *Review of Educational Research,* 52**,** 201-217.
HORNE, J., BEJTKA, K. & MILLER, S. 2008. *The Trends in International Maths and Science Survey (TIMSS) is an international assessment of pupil attainment in maths and science at primary and secondary school level.* [Online]. Available: http://www.scotland.gov.uk/Publications/2009/10/13150724/0.
ILTUS, S. 2006 Significance of home environments as proxy indicators for early childhood care and education. *Paper commissioned for the EFA Global Monitoring Report 2007, Strong foundations: early childhood care & education.* UNESCO.
JACKMAN, S. 2009. *Markov Chain Monte Carlo,* Chichester, John Wiley& Sons Ltd.
JAEGER, T. F. 2008. Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of memory and language,* 59**,** 434-446.
JAWORSKI, B. 1993. The Professional Development of Teachers — The Potential of Critical Reflection. *British Journal of In-Service Education,* 19**,** 37-42.

JIANG, J. 2010. Large sample techniques for statistics. New York: Springer.
JIMÉNEZ-ALEIXANDRE, M. P. & ERDURAN, S. 2007. Argumentation in Science Education: An Overview. *Contemporary Trends and Issues in Science Education,* 35**,** 3-27.
KASS, R. & RAFTERY, A. 1995. Bayes Factors. *Journal of the American Statistical Association,* 90**,** 773-795.
KASS, R. E., CARLIN, B. P., GELMAN, A. & NEAL, R. M. 1998. Markov Chain Monte Carlo in Practice: A Roundtable Discussion. *The American Statistician,* 52**,** 93-100.
KATHIE ANDERSON, RHONDA BROOKS & SR MOLLIE REAVIS. PROJECT FIFTEEN: INSTRUCTIONAL STRATEGIES TO ACHIEVE GENDER EQUITY. Available: http://www.prenhall.com/divisions/esm/app/ph-elem/multicult/html/chap15.html [Accessed 1st June 2014].
KILPATRICK, J. 1987. What constructivism might be in mathematics education. *In:* BERGERON, J. C., HERSCOVICS, N. & KIERAN, C. (eds.) *Proceedings of the 11th PME International Conference.*
KIRSCH, I. S., OECD & PISA 2002. *Reading for change : performance and engagement across countries : results from PISA 2000 / Irwin Kirsch ... [et al.],* Paris :, Organisation for Economic Co-operation and Development.
KUHA, J. 2004. AIC and BIC: Comparisons of Assumptions and Performance. *Sociological Methods & Research,* 33**,** 188-229.
KUHN, D. 2010. Teaching and learning science as argument. *Science Education,* 94**,** 810-824.
KUHN, D. & UDELL, W. 2003. The Development of Argument Skills. *Child Development,* 74**,** 1245-1260.
LABOVITZ, S. 1967. Some Observations on Measurement and Statistics. *Social Forces,* 46**,** 151-160.
LAVE, J. 1988. *Cognition in Practice: Mind,mathematics, and culture in everyday life,* Cambridge, UK, Cambridge University Press.
LEARNING AND TEACHING SCOTLAND 2004. *Improving science education 5-14 : 5-14 teacher support / Learning and Teaching Scotland,* Dundee, Learning and Teaching Scotland.
LEARNING AND TEACHING SCOTLAND. 2006. *Assessment is for Learning: self-assessment toolkit* [Online]. Available: http://www.wiredshire.org.uk/professional/support/csg/english/documents/AifLToolkitforschools.pdf.
LEBANON, G. 2006. Metropolis-Hastings and Gibbs Sampling.
LEITHWOOD, K., HARRIS, A. & HOPKINS, D. 2008. Seven strong claims about successful school leadership. *School Leadership & Management: Formerly School Organisation,* 28**,** 27 - 42.
LEVIN, B. 2004. Making Research Matter More. *Education Policy Analysis Archives,* 12.
LEVIN, B. 2013. To know is not enough: research knowledge and its use. *Review of Education,* 1**,** 2-31.
LIEBECK, P. 1984. *How children Learn Mathematics,* Middlesex, Penguin.
LIPSEY, M., PUZIO, K., YUN, C., HERBERT, M., STEINKA-FRY, K., COLE, M., ROBERTS, M., ANTHONU, K. & BUSICK, M. 2012. *Tranlating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms,* Washington, National Centre for Special Education Rsearch.
LITTLE, R. J. A. & RUBIN, D. B. 1987. *Statistical analysis with missing data,* New York, Wiley.
LONDON MATHEMATICAL SOCIETY 1995. *Tackling the Mathematics Problem,* London, LMS.
LORD, F. M. 1962. Estimating norms by item-sampling *Educational and Psychological Measurement,* 34**,** 259-299.

LORD, F. M. 1977. Practical Applications of Item Characteristic Curve Theory. *Journal of Educational Measurement,* 14**,** 117-138.
LUNETTA, V. N., HOFSTEIN, A. & CLOUGH, M. P. 2007. Teaching and learning in the school science laboratory. An analysis of research, theory, and practice. *In:* ABELL, S. K. & LEDERMAN, N. G. (eds.) *Handbook of research on science education.* Mahwah,NJ: Lawrence Erlbaum Associates.
MACINTYRE, T. & FORBES, I. 2002. Algebraic Skills and CAS - Could Asessment Sabotage the Potential? *The International Journal of Computer Algebra in Mathematics Education,* 9**,** 29-56.
MANSELL, W., JAMES, M. & ASSESSMENT REFORM GROUP 2009. Assessment in schools. Fit for purpose? A Commentary by the Teaching and Learning Research Programme. *Teaching and Learning Research Programme.* London.
MASON, J. & SPENCE, M. 1999. Beyond mere knowledge of mathematics: The importance of knowing-to act in the moment. *Educational Studies in Mathematics,* 38**,** 135-161.
MCNAB, D. 1999. Mathematics education in Scottish schools: an uncertain vision? *Scottish Educational Review* [Online], 31. Available: http://www.scotedreview.org.uk/pdf/62.pdf.
MERCER, N. 2008. *Three kinds of talk* [Online]. Available: https://thinkingtogether.educ.cam.ac.uk/resources/5_examples_of_talk_in_groups.pdf.
MERCER, N., DAWES, L., WEGERIF, R. & SAMS, C. 2004. Reasoning as a Scientist: Ways of Helping Children to Use Language to Learn Science. *British Educational Research Journal,* 30**,** 359-377.
MERCER, N. & SAMS, C. 2006. Teaching Children How to Use Language to Solve Maths Problems. *Language and Education,* 20**,** 507-528.
MERCER, N., WEGERIF, R. & DAWES, L. 1999. Children's Talk and the Development of Reasoning in the Classroom. *British Educational Research Journal,* 25**,** 95-111.
MICHELL, J. 1986. Measurement Scales and Statistics: A Clash of Paradigms. *Psychological Bulletin,* 100**,** 398-407.
MILLAR, R. 1989. Constructive criticisms. *International Journal of Science Education,* 11**,** 587-596.
MISLEVY, R. J. 1984. Estimating latent distributions *Psychometrika,* 49**,** 359-381.
MISLEVY, R. J. 1985. Estimation of Latent Group Effects. *Journal of the American Statistical Association,* 80**,** 993-997.
MISLEVY, R. J., BEATON, A. E., KAPLAN, B. & SHEEHAN, K. M. 1992a. Estimating Population Characteristics from Sparse Matrix Samples of Item Responses. *Journal of Educational Measurement* 29**,** 133-161.
MISLEVY, R. J., JOHNSON, E. G. & MURAKI, E. 1992b. Scaling Procedures in NAEP. *Journal of Educational Statistics,* 17**,** 131-154.
MORROW, C. & MORROW, J. 2005. Connecting Women with Mathematics. *In:* KAISER, G., ROGERS, D. B. P. H. P. & ROGERS, P. (eds.) *Equity In Mathematics Education: Influences Of Feminism And Culture.* London: Taylor & Francis.
MORTIMER, E. F. & SCOTT, P. H. 2003. *Meaning Making in Secondary Science Classrooms,* Maidenhead, OUP.
MUELLER, C. W. & PARCEL, T. L. 1981. Measures of Socioeconomic Status: Alternatives and Recommendations. *Child Development,* 52**,** 13-30.
MULLIS, I. V. S., MARTIN, M. O. & FOY, P. 2008. *TIMSS 2007 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 Centre, Boston College.

NATIONAL COUNCIL OF TEACHERS OF MATHEMATICS (NCTM) 1989. *Curriculum and Evaluation Standrads for School Mathematics,* Virginia, USA, NCTM.
NATIONAL COUNCIL OF TEACHERS OF MATHEMATICS (NCTM) 1991. *Professional Standards for Teaching Mathematics,* Virginia, USA, NCTM.
NATIONAL GOVERNORS ASSOCIATION (NGA). 2010. *Common Core State Standards Initiative* [Online]. Available: http://www.corestandards.org/resources.
NIEMI, H. 2002. Active learning—a cultural change needed in teacher education and schools. *Teaching and Teacher Education,* 18**,** 763-780.
NORWEGIAN MINISTRY OF EDUCATION AND RESEARCH 2010. Science for the Future: Strategy for Strengthening Mathematics, Science and Technology (MST) 2010–2014. Oslo: Norwegian Ministry of Education and Research.
NOVAK, J. D. 1986. The importance of emerging constructivist epistemology for mathematics education. *Journal of Mathematical Behaviour,* 5**,** 181-184.
NUSSBRAUM, J. & NOVICK, S. 1981. Brainstorming in the classroom to invent a model: A case study *School Science Review,* 62**,** 771-778.
NUTLEY, S., DAVIS, H. & WALTER, I. 2003. Evidence-Based Policy and Practice: Cross-Sector Lessons from the United Kingdom. *Social Policy Journal of New Zealand,* 20.
OANCEA, A. & PRING, R. 2009. The importance of being thorough: On systematic accumulations of "what works" in education research. *In:* BRIDGES, D., SMEYERS, P. & SMITH, R. (eds.) *Evidence-Based Education Policy: What Evidence? What Basis? Whose Policy?* Oxford: Blackwells.
ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD), 2013. *Programme for International Student Assessment (PISA) results from 2012: United Kingdom Country Note*. Available: http://www.oecd.org/unitedkingdom/PISA-2012-results-UK.pdf
OSBORNE, J. 1996. Beyond Constructivism. *Science Education,* 80**,** 53-82.
OSBORNE, J. & HENNESSY, S. 2003. Literature Review in Science Education and the Role of ICT: Promise, Problems and Future Directions. Bristol: Futrurelab.
PAGE-BUCCI, H. 2003. *The value of Likert scales in measuring attitudes of online learners* [Online]. Available: www.hkadesigns.co.uk/websites/msc/reme/likert.htm.
PEHKONEN, E., AHTEE, M. & LAVONEN, J. 2007. *How Finns Learn Mathematics and Science,* Rotterdam, Sense Publishers.
PHILLIPS, D. C. 1995. The Good, the Bad, and the Ugly: The Many Faces of Constructivism. *Educational Researcher,* 24**,** 5-12.
POWNEY, J. 1996. Gender and Attainment. Edinburgh: Scottish Council for Research in Education (SCRE).
RAFTERY, A. E. & LEWIS, S. M. 1992. [Practical Markov Chain Monte Carlo]: Comment: One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo. *Statistical Science,* 7**,** 493-497.
ROBERTS, G. 2002. SET for success: The supply of people with science, technology, engineering and mathematical skills.
ROGOFF, B. 1994. Developing understanding of the idea of communities of learners. *Mind, Culture, and Activity,* 1**,** 209-229.
ROGOFF, B., PARADISE, R., ARAUZ, R. M., CORREA-CHÁVEZ, M. & ANGELILLO, C. 2003. Firsthand Learning Through Intent Participation. *Annual Review of Psychology,* 54**,** 175-203.
ROSSI, P., ALLENBY, G. M. & MCCULLOCH, R. 2005. *Bayesian Statistics and Marketing,* West Sussex, John Wiley & Sons Ltd.
RUBIN, D. B. 1987. *Multiple imputation for nonresponse in surveys,* Chichester, Wiley.

RUST, K. 1985. Variance Estimnation fro Complex Estimators in Sample Surveys. *Journal of Official Statistics,* 1**,** 381-397.
RUTKOWSKI, L. & RUTKOWSKI, D. 2010. Getting it 'better': the importance of improving background questionnaires in international large-scale assessment. *Journal of Curriculum Studies,* 42**,** 411 - 430.
SANDERSON, I. 2003. Is it 'what works' that matters? Evaluation and evidence-based policy-making. *Research Papers in Education,* 18**,** 331-345.
SCHAFER, J. L. 1999. Multiple imputation: a primer. *Statistical Methods in Medical Research,* 8**,** 3-15.
SCHAFER, J. L. 2003. Multiple Imputation in Multivariate Problems When the Imputation and Analysis Models Differ. *Statistica Neerlandica,* 75**,** 19-35.
SCHAGEN, I. & ELLIOT, K. 2004. *But what does it mean? The use of effect sizes in educational research,* Berks, NFER.
SCHIFTER, D. 1996a. A constructivist perspective on teaching and learning of mathematics. *In:* FOSNOT, C. T. (ed.) *Constructivism: Theory, perspectives and practice.* New York: Teachers College Press.
SCHIFTER, D. 1996b. On teaching and learning mathematics. *Phi Delta Kappan,* 77**,** 492.
SCHOENFELD, A. H. 2002. Making Mathematics Work for All Children: Issues of Standards, Testing, and Equity. *Educational Researcher,* 31**,** 13-25.
SCHOENFELD, A. H. 2004. The Maths Wars. *Educational Policy,* 18**,** 253-286.
SCHOOL CURRICULUM AND ASSESSMENT AUTHORITY (SCAA) 1994. National Curriculum.
SCIENCE AND ENGINEERING EDUCATION ADVISORY GROUP (SEEAG). 2012. *Supporting Scotland's STEM Education and Culture* [Online]. Available: http://www.scotland.gov.uk/Publications/2012/02/4589/0.
SCIENCE COMMUNITY REPRESENTING EDUCATION (SCORE) 2008. Practical work in science: a report and proposal for a strategic framework. London: SCORE.
SCIENCE COMMUNITY REPRESENTING EDUCATION (SCORE) 2009. Explore, inspire, discover: practical work in science. London: SCORE.
SCOTTISH EXECUTIVE 1999. Standards and Quality in Secondary Schools 1995-1999: Mathematics. Edinburgh: Scottish Executive.
SCOTTISH EXECUTIVE. 2004a. *Ambitious, Excellent Schools - Our Agenda for Action* [Online]. Available: http://www.scotland.gov.uk/Publications/2004/11/20176/45852.
SCOTTISH EXECUTIVE 2004b. Curriculum for Excellence: Ministerial Response. Edinburgh: SE.
SCOTTISH EXECUTIVE 2004c. Curriculum for Excellence: The Curriculum Review Group. Edinburgh: SE.
SCOTTISH GOVERNMENT 2007. *Government Economic Strategy*.
SCOTTISH GOVERNMENT 2009a. Assessment for Curriculum for Excellence: strategic vision and key principles.
SCOTTISH GOVERNMENT. 2009b. *Creative Industries Key Sector Report* [Online]. Available: http://www.scotland.gov.uk/Publications/2009/11/24133819/0.
SCOTTISH GOVERNMENT. 2009c. *Scottish Survey of Achievement (SSA) 2008: Mathematics and Core Skills. Technical Annex* [Online]. Available: http://www.scotland.gov.uk/Publications/2009/08/12134459/0.
SCOTTISH GOVERNMENT. 2011a. *Building the Curriculum 5: A framework for assessment* [Online]. Available: http://www.educationscotland.gov.uk/Images/BtC5Framework_tcm4-653230.pdf.
SCOTTISH GOVERNMENT. 2011b. *Excellence in Mathematics: Report from the Maths Excellence Group* [Online]. Available: http://www.scotland.gov.uk/Resource/Doc/91982/0114466.pdf.

SCOTTISH GOVERNMENT 2011c. Government Economic Strategy.
SCOTTISH OFFICE EDUCATION AND INDUSTRY DEPARTMENT (SOEID) 1996a. *Achievement for all : a report on selection within schools / by HM inspectors of schools,* Edinburgh, HMSO.
SCOTTISH OFFICE EDUCATION AND INDUSTRY DEPARTMENT (SOEID) 1996b. *Standards and quality in Scottish schools, 1992-95 : a report / by HM Inspectors of Schools, Audit Unit,* Edinburgh, Scottish Office.
SCOTTISH OFFICE EDUCATION AND INDUSTRY DEPARTMENT (SOEID) 1997a. *Achieving success in S1/S2 : a report on the review of provision in S1/S2 by HM Inspectors of Schools,* Edinburgh, SOEID.
SCOTTISH OFFICE EDUCATION AND INDUSTRY DEPARTMENT (SOEID) 1997b. *Improving mathematics education 5-14 : a report / by HM Inspectors of Schools,* Edinburgh, Stationery Office.
SCOTTISH OFFICE EDUCATION DEPARTMENT (SOED) 1991. Curriculum and Assessment in Scotland National Guidelines: Mathematics 5-14. Edinburgh: SOED.
SCOTTISH OFFICE EDUCATION DEPARTMENT (SOED) 1993. *Effective learning and teaching in Scottish secondary schools: mathematics : a report / by HM Inspectors of Schools,* Edinburgh, Scottish Office Education Department.
SCOTTISH OFFICE EDUCATION DEPARTMENT (SOED) 2000. Environmental Studies - Society, Science and Technology: 5-14 National Guidelines. Edinburgh: SOED.
SCOTTISH QUALIFICATION AUTHORITY (SQA) 1999. Standard Grade Arrangements in Mathematics. Glasgow: Scottish Qualification Authority.
SHOEMAKER 1973. *Principles and Procedures of Multiple Matrix Sampling,* Cambridge, MA, Ballinger Publishing Company.
SHULMAN, L. S. & TAMIR, P. 1973. Research on teaching in the natural sciences. *In:* TRAVERS, R. M. V. (ed.) *Second handbook of research on teaching.* Chicago: Rand McNally.
SIERPINSKA, A. 1994. *Understanding in Mathematics,* London, Falmer Press.
SIMPSON, T. L. 2002. Dare I Oppose Constructivist Theory? *The Educational Forum,* 66**,** 347-354.
SINHARAY, S., STERN, H. & RUSSELL, D. 2001. The Use of Multiple Imputation for the Analysis of Missing Data. *Psychological Methods,* 6**,** 317-329.
SIRIN, S. R. 2005. Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research,* 75**,** 417-453.
SIROTNIK, K. & WELLINGTON, R. 1977. Incidence Sampling: An Integrated Theory for "Matrix Sampling". *Journal of Educational Measurement,* 14**,** 343-399.
SKEMP, R. R. 1987. *Psychology of Learning Mathematics,* Hillsdale, N.J. ; Hove :, L. Erlbaum Associates.
SNELL, E. J. 1964. A Scaling Procedure for Ordered Categorical Data *Biometrics,* 20**,** 592-607.
SOLESBURY, W. & ESRC 2001. Evidence based policy: Whence it came and where it's going. ESRC UK Centre for Evidence Based Policy and Practice London.
SOLOMON, J. 1994. The Rise and Fall of Constructivism. *Studies in Science Education,* 23**,** 1-19.
SOLOMON, J. 2002. Science Stories and Science Texts: What can they do for our students? *Studies in Science Education,* 37**,** 85-105.
SPELKE, E. 2005. Sex Differences in Intrinsic Aptitude for mathematics and Science? *American Psychologist,* 60**,** 950-958.
SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. & LINDE, A. V. D. 2002. Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology),* 64**,** 583-639.

STEFFE, L. P. 1995. Alternative epistemologies: An educator's perspective. *In:* STEFFE, L. & GALE, J. (eds.) *Constructivism in Education.* Hillsdale, NJ: Erlbaum.
STEVENS, S. S. 1946. On the Theory of Scales of Measurement. *Science,* 103**,** 677-680.
STURMAN, L., RUDDOCK, G., BURGE, B., STYLES, B., LIN, Y. & VAPPULA, H. 2008. *England's Achievement in TIMSS 2007: National Report for England* [Online]. Slough: NFER.
SWAN, M. 2006. Designing and using research instruments to describe the beliefs and practices of mathematics teachers. *Research in Education***,** 58-127.
TABER, K. S. 2011. Guiding the practice of constructivist teaching. *Teacher Development,* 15**,** 117-122.
TINKLIN, T., CROXFORD, L., DUCKLIN, A. & FRAME, B. 2001. Gender and Pupil Performance in Scotland's Schools. *In:* SCOTTISH EXECUTIVE EDUCATION DEPARTMENT (ed.). Edinburgh: SEED.
TINSLEY, H. & TINSLEY, D. 1987. Uses of Factor Analysis in Counseling Psychology Research. *Journal of Counseling Psychology,* 34**,** 414-424.
TOWNSEND, J. T. & ASHBY, F. G. 1984. Measurement Scales and Statistics: The Misconception Misconceived. *Psychological Bulletin,* 96**,** 394-401.
TYMMS, P. 2004. Effect sizes in multilevel models. *In:* SCHAGEN, I. & ELLIOT, K. (eds.) *But what does it mean?The use of effect sizes in educational research.* Slough: NFER.
TYMMS, P., MERRELL, C. & HENDERSON, B. 1997. The First Year at School: A Quantitative Investigation of the Attainment and Progress of Pupils∗. *Educational Research and Evaluation,* 3**,** 101-118.
VAN DEN HEUVEL-PANHUIZEN, M. 2001. Realistic Mathematics Education in the Netherlands. *In:* ANGHILERI, J. (ed.) *Principles and practice in arithmetic teaching.* Buckingham/Philadelphia: Open University Press.
VAN DEN HEUVEL-PANHUIZEN, M. 2005. The role of contexts in assessment of problems in mathematics. *For the learning of Mathematics* 25, no. 2:2-9
VAN HIELE, P. 1999. Developing Geometric Thinking Through Activities that Begin with Play. *Teaching Children Mathematics***,** 310-316.
VON DAVIER, M., GONZALEZ, E. & MISLEVY, R. J. 2009. What are plausible values and why are they useful? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments* 2**,** 9-36.
VON GLASERSFELD, E. 1989. Constructivism in Education. *In:* HUSEN, T. & POSTLETHWAITE, T. N. (eds.) *The International Encyclopaedia of Education,Supplement Vol 1.* New York: Pergamon Press.
VOSNIADOU, S. 1996. Towards a revised cognitive psychology for new advances in learning and instruction. *Learning and Instruction,* 6**,** 95-109.
VYGOTSKY, L. S. 1978. *Mind in society:The development of higher psychological processes,* Cambridge, MA, HarvardUniversity Press.
VYGOTSKY, L. S. 1991. Genesis of the higher mental functions. *In:* LIGHT, P., SHELDON, S. & WOODHEAD, M. (eds.) *Learning to think : a reader* London: Routledge in assoc. with Open Univ. P.
VYGOTSKY, L. S. & KOZULIN, A. 1986. Thought and language. Cambridge, Mass.: MIT Press.
WARING, S. 2000. *Can You Prove It? Developing Concepts of Proof in Primary and Secondary Schools*, Mathematical Association.
WATSON, A. 2008. How secondary teachers structure the subject matter of mathematics. *British Society for Research into Learning Mathematics Proceedings,* 28.

WELLS, G. & ARAUZ, R. M. 2006. Dialogue in the Classroom. *Journal of the Learning Sciences,* 15**,** 379-428.

WENGER, E. 2006. *Communities of Practice* [Online]. Available: wenger-trayner.com/theory/ [Accessed October 21 2013].
WHITE, R. & GUNSTONE, R. 1992. *Probing Understanding,* London, The Falmer Press.
WHITE, R. T. 1979. Relevance of practical work to comprehension of physics. *Physics Education,* 14**,** 384.
WHITEHEAD, A. N. 1929. *The Aims of Education and Other Essays,* New York, Macmillan.
YAIR, G. 2000. Educational Battlefields in America: The Tug-of-War over Students' Engagement with Instruction. *Sociology of Education,* 73**,** 247-269.
YOONG, W. K. 1987. Aspects of Mathematical Understanding. *Singapore Journal of Learning Disabilities,* 37.
YOUNG, K., ASHBY, D., BOAZ, A. & GRAYSON, L. 2002. Social Science and the Evidence-based Policy Movement. *Social Policy and Society,* 1**,** 215-224.

Appendix 3. Methodological Issues

Appendix 3.1. **Buffon's Needle Experiment**

ON AN EXPERIMENTAL DETERMINATION OF π

ASAPH HALL

In his *Theorie analytique des Probabilités*, Chap. V., Laplace has shown that we may make use of this calculus to determine the lengths of curves and to find their surfaces; and he has pointed out very briefly how this may be done. Imagine a plane on which are drawn equidistant and parallel right lines, and let there be thrown on this plane at random a right line of given length. It is required to find the probability that the right line will intersect one of the parallel lines. This is one of the questions solved by Laplace, and by varying his solution a little, it is easy to find that its probability is expressed by the definite integral

$$\int_0^{\frac{1}{2}\pi} \frac{2l}{a\pi} \cos\phi d\phi = \frac{2l}{a\pi};$$

where a is the interval of the parallel lines, and l is the length of the random line. If we denote by m the whole number of times the line is thrown on the plane, and by n the number of intersections, and if m be very great and the trials be made so that there is no systematic error in the experiments, we may assume that the probability is expressed by the ratio $\frac{n}{m}$. Equating this to the rigorous value, we have

$$\pi = \frac{2ml}{an} \tag{1}$$

In this expression a and l are known, and m and n are to be found by observation.

In 1864, my friend Capt. O. C. Fox was unable to do active duty on account of a severe wound, and I proposed that he should make some experiments for determining the ratio $\frac{n}{m}$. Capt. Fox had made a plane wooden surface ruled with equidistant parallel lines, and on this he threw at random a fine steel wire. After making the first set of experiments, and in order to avoid as much as possible any constant error that might arise from his position or manner of holding the rod over the surface, the surface was given a slight rotatory motion before dropping the rod; the following are the results of the experiments of Capt. Fox:

m	n	l	a	
500	236	3 inches	4 inches	surface stationary.
530	253	3 inches	4 inches	surface revolved
590	939	5 inches	2 inches	surface revolved.

Substituting these numbers in formula (1), we have

$$\pi = \frac{2 \cdot 500 \cdot 3}{4 \cdot 236} = 3.1780,$$
$$\pi = \frac{2 \cdot 530 \cdot 3}{4 \cdot 253} = 3.1423,$$
$$\pi = \frac{2 \cdot 590 \cdot 5}{2 \cdot 939} = 3.1416.$$

Washington, *June 5*, 1872

Source: (http://www.cs.xu.edu/math/Sources/Buffon/Hall%20on%20Buffon%20Needle.pdf)

Appendix 3.2. Combining multiple imputations using Rubin's rules

When performing a multiply-imputed analysis the results from each of *m* imputed data sets are combined to obtain a set of results. From each analysis, one must first compute and save the estimates and standard errors. Suppose that $\hat{Q}_j$ is an estimate of a scalar quantity of interest obtained from data set *j* (*j* =1, 2… *m*) and $\hat{U}_j$ is the squared standard error associated with $\hat{Q}_j$. These statistics are then combined to get:

i. The overall point estimate, which is just the average of the individual estimates

$$\bar{Q} = \frac{1}{m}\sum_{j=1}^{m} \hat{Q}_j$$

ii. The total variance (and thereafter the standard error to compute confidence intervals), which is calculated by combining:

   a. the *within-imputation* variance

$$\bar{U} = \frac{1}{m}\sum_{j=1}^{m} \hat{U}_j$$

   and

   b. the *between-imputation* variance

$$B = \frac{1}{m-1}\sum_{j=1}^{m} (\hat{Q}_j - \bar{Q})^2$$

   to give

   c. the total variance

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B$$

Appendix 3.3. **Efficiency of using *m* imputations**

Efficiency of any estimate is based on the extent of missing data. The variation in results across imputed data sets is what reflects the statistical uncertainty attributed to missing data. The relative increase in variance due to non-response is given by $\mathrm{r} = \frac{(1+\mathrm{m}^{-1})\mathrm{B}}{\bar{\mathrm{U}}}$, the ratio of between-imputation to within-imputation contribution to total variance.

The estimated rate of missing information, based on r, is approximately $\lambda = \frac{\mathrm{r}}{1+\mathrm{r}}$. A more refined estimate of this fraction, obtained by comparing the spread of the variance components detailed in Schafer (1999: 5) is:

$$\lambda = \frac{r + 2/(df + 3)}{r + 1}$$

Many are surprised by Rubin's claim that only 3-10 imputations are needed to achieve highly efficient estimates when working with multiple imputations. Rubin (1987: 114) shows that the efficiency of an estimate based on m imputations is approximately

$$\left(1 + \frac{\lambda}{\mathrm{m}}\right)^{-1}$$

Rubin presents the following efficiencies for various values of *m* and rates of missing information, with Schafer's λ being used in place of Rubin's γ for consistency with arguments presented above.

	λ				
m	0.1	0.3	0.5	0.7	0.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

These data show that in most situations only a small number of imputations are required, unless the rate of missing information is particularly high. Rubin's recommendation is to use five imputations unless the empirical data suggests otherwise.

Appendix 3.4. **Rubin's Rules**

The procedures for combining data using Rubin's rules for imputed data are set out in Chapter 3 but here the practical process is presented by way of explaining tables of results in subsequent chapters. First the four key data points for each plausible value are computed from each multilevel model – the coefficient for the explanatory variable, the associated standard error, effect size and error for effect. The tabulated values in Table 3.4-1 include the computation outlined above for level of significance, making five entries for each plausible value.

Table 3.4-1: Column headings for 1st PV

B	C	D	E	F
COEFFICIENT BASED ON BMATHS_PV01_MCMC	S.E.	Sig	effect size	+/- error

Those five values for each plausible value are concatenated to generate a long row for each response category, with five sets of five entries. The first entry in each group is placed in columns B, G, L, Q and V; populating groups of five columns with headings that relate to the calculations set out in Table 3.4-2. The separate elements are all straight forward computations once the data is set up in the correct structure, providing the overall estimate $\bar{Q}$ and overall SE copied for onward transmission to the findings in the models using all five plausible values.

Table 3.4-2: Calculations for combining imputed data (Rubin's Rules)

Overall estimate (Q bar, $\bar{Q}$)	Within Imputation Variance (U bar, $\bar{U}$)	Between Imputation Variance (B)	Total Variance (T)	Overall SE
=AVERAGE (B1,G1,L1, Q1,V1)	=AVERAGE ($C1^2$,$H1^2$,$M1^2$,$R1^2$,$W1^2$)	=VAR.S (B1,G1,L1, Q1,V1)	$=\bar{U}+\left(1+\frac{1}{5}\right)B$	=SQRT(T)

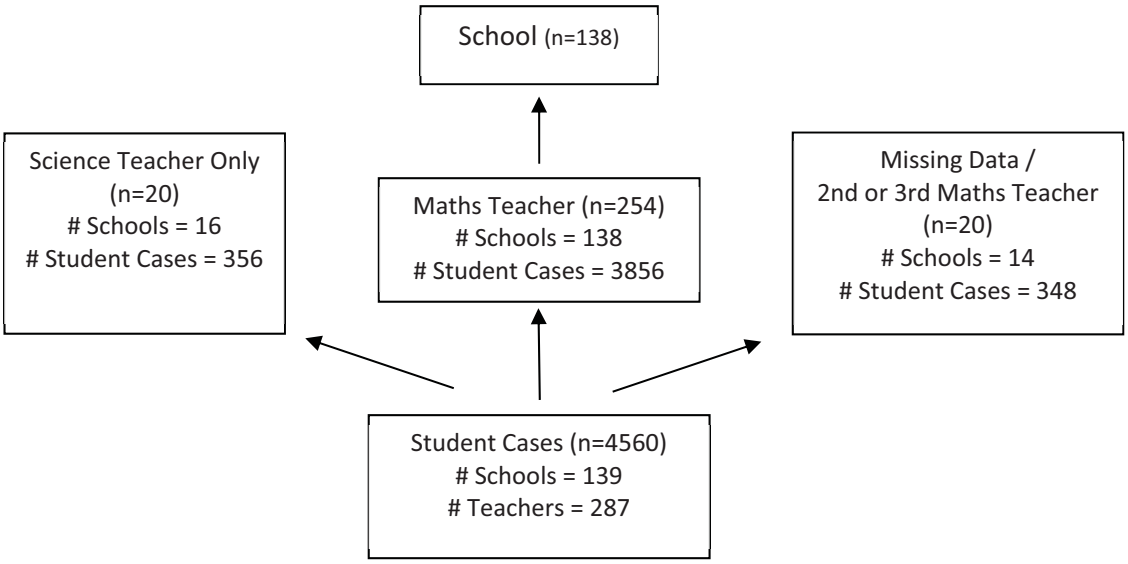
A fully illustrated example of this process in practice is presented in Chapter 5; other models discussed in the findings will only extract the relevant two entries for the overall coefficient and overall standard error that permit calculation of effect size, error bounds and standard Normal equivalent to contribute to discussion.

Appendix 3.5. **Data structure for G4 Maths and Science**

A first step in managing the data sought to reduce the data so that a unique set of cases was created for mathematics and science experiences for separate analysis. To simplify the model to have a G4 Mathematics data set the student cases with teachers who do not teach mathematics (i.e. science only) were dropped and duplicate student cases with more than one mathematics teacher were also dropped. Other cases where the *teacher* or *school* data are missing (unit non-response) were removed since those cases would not contribute to any hierarchical model that drew on those levels of the hierarchy. This identified 3856 cases to be retained for analysis, as shown in Figure 3.5-1and Figure 3.5-2, with each case representing a student's cognitive and background data alongside one mathematics teacher's background data and associated school data. This reduction does not represent a particularly significant loss in potential data because there were originally 3929 students making up the 4560 student cases; a loss of about 1% of students originally sampled.

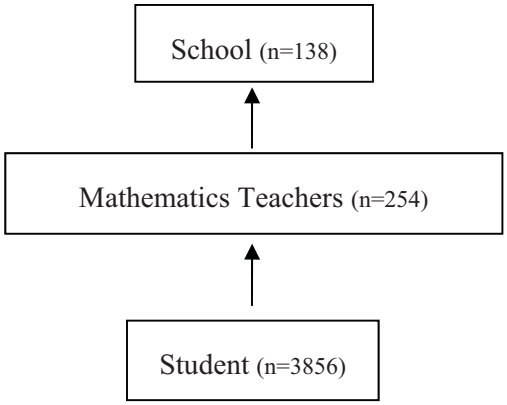
To decide which case to retain when there was a second (or third) mathematics teacher, I reviewed the responses to the teachers' background questionnaire data that addressed teacher's age and years of experience in the classroom, both regarded as potential explanatory variables in a subsequent model. I sought to balance those data across the retained and deleted cases. The resultant averages for teaching experience in the retained group was 12.7 years and for the deleted cases it was 14.6 years; both lying within the same band of 'years of teaching experience'.

Figure 3.5-1: Student-Teacher structure for Mathematics



Finally, there was one teacher working with two classes. Given the atomic unit is at student level these two classes were kept in the data set, the equivalent of combining the classes. The final simplified structure and composition of the data set for analysis of mathematics related achievement is presented in Figure 3.5-2.

Figure 3.5-2: Simplified Nested Hierarchical Classification (Gr4 Mathematics)



A similar classification was carried out to prepare the data set for the analysis of science achievement that took into consideration the science teacher's background data; and similarly for the handling of G8 data on mathematics and science education.

Appendix 3.6. **Effect size**

Effect sizes as outlined in Chapter 3 are readily computed using appropriate formulae in Excel. Illustrative content is presented in Table 3.6-1 that includes the calculations for error bounds on effect sizes.

Table 3.6-1: Excel formulae for Effect Size and error bounds

	A	B	C	H	I	K	L
1	Response	GR4_ MATHS _NULL	S.E.	GR4_MATH S_EXPLAIN ASMMATPV	S.E. (SIG)	effect size =H6/ SQRT(B$41)	+/- error =SQRT((I6/H6)^2 +(I$41/H$41)^2)*K6
5	EXPLAIN SOME LESSONS			13.1	3.4 (**)	0.20 VPC	0.05
37	Level: IDSCHOOL	408.3	162.7	53.9	70.5	0.01	
39	Level: IDTEACH	740.5	171.8	518.9	104.8	0.12	
41	Level: IDSTUD	4130.3	105.0	3204.1	82.3	0.71	
43	Level: IDCASE	704.7	8.3	704.7	8.3	0.16	

Significance is determined by comparing ratio of coefficient with standard error, through an expansion of an If, Condition, Then, Else statement:

$$= \text{IF}(\text{ABS}(\text{H6/I6}) > 2.58, " ** ", \text{IF}(\text{ABS}(\text{H6/I6}) > 1.96, " * ", " - "))$$

Effect size is evaluated using $\Delta = \frac{\beta_1}{\sigma_e}$ or $\Delta = \frac{2\beta_1}{\sigma_e}$ dependent on categorical or continuous response, through a compute function using student-level variance (σ_e) from the null model:

$$= H6/SQRT(B\$41)$$

Error bounds are similarly evaluated using a compute function with the student-level variance and standard error:

$$= SQRT((I6/H6)\text{^}2 + (C\$41/B\$41)\text{^}2) * K6$$

The standard Normal distribution equivalent for any given effect size comes from an inbuilt Excel function:

$$= NORM.S.DIST(K6, TRUE)$$

In this case (Table 3.6-1) the equivalent proportion is 58%, i.e. the effect size of 0.20 for 'explaining answers in some lessons' equates to those students exceeding the scores of 58% of the control group.

Appendix 4. Additional Tables from EDA

Appendix 4.1. Background Data on G4 Students

Table 4.1-1: About how many books are there in your home?

NUM BOOKS AT HOME		Freq.	Percent	Cumulative Percent
Valid	NONE OR VERY FEW (0 TO 10 BOOKS)	6,500	12.1	12.1
	ONE SHELF (11 TO 25 BOOKS)	10,500	19.7	31.8
	ONE BOOKCASE (26 TO 100 BOOKS)	17,500	32.8	64.6
	TWO BOOKCASES (101 TO 200 BOOKS)	9,900	18.6	83.2
	THREE OR MORE BOOKCASES (OVER 200 BOOKS)	9,000	16.8	100.0
	Total	53,400	100.0	

Figure 4.1-1: Average of achievement– mean score (95% CI) by Study Tools

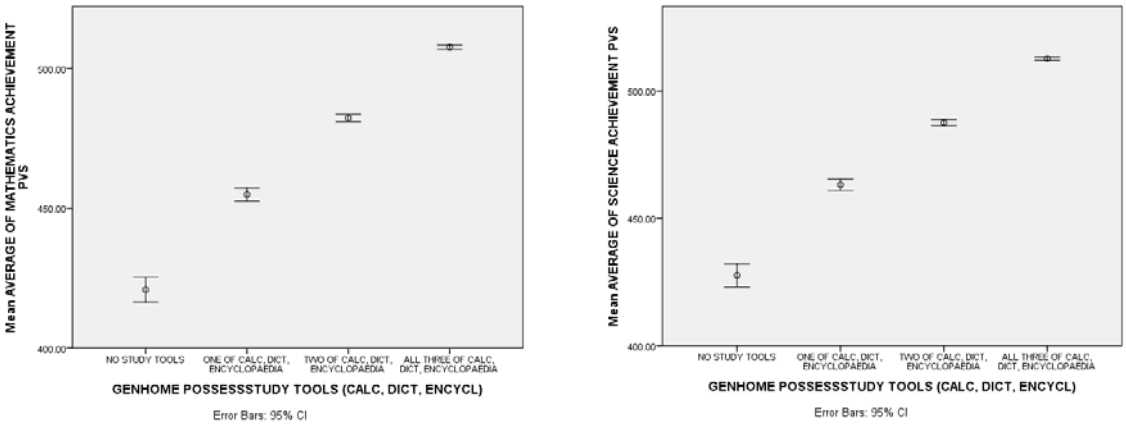


Table 4.1-2: Index for Social activity (ASDSOC) - distribution and associated achievement (s.d.)

		Frequency	Percent	Cumulative Percent	Mean Mathematics Achievement	Mean Science Achievement
Valid	LOW	10,100	18.8	18.8	481.9 (81.5)	493.1 (77.4)
	MID	25,500	47.5	66.3	502.4 (72.7)	508.2 (69.5)
	HIGH	18,100	33.7	100.0	491.2 (73.6)	493.6 (69.4)
	Total	53,600	100.0		494.8 (75.1)	500.4 (71.4)

Table 4.1-3: Index for Personal Development (ASDPDEV) - distribution and associated achievement (s.d.)

		Frequency	Percent	Cumulative Percent	Mean Mathematics Achievement	Mean Science Achievement
Valid	LOW	28,400	52.9	52.9	496.1 (72.6)	499.9 (66.8)
	MID	23,400	43.5	96.5	498.0 (75.9)	505.6 (74.0)
	HIGH	1,900	3.5	100.0	437.4 (78.0)	445.0 (79.7)
	Total	53,600	100.0		494.8 (75.1)	500.5 (71.4)

Figure 4.1-2: Mean achievement scores by index of proportion individual

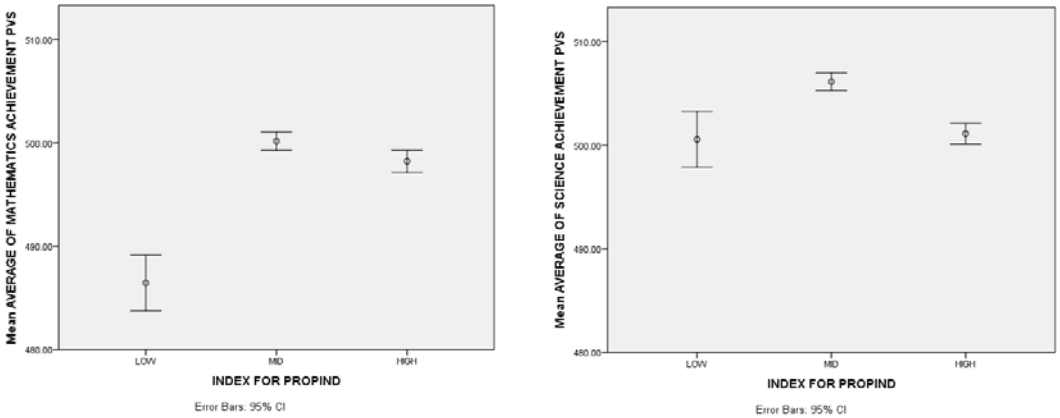
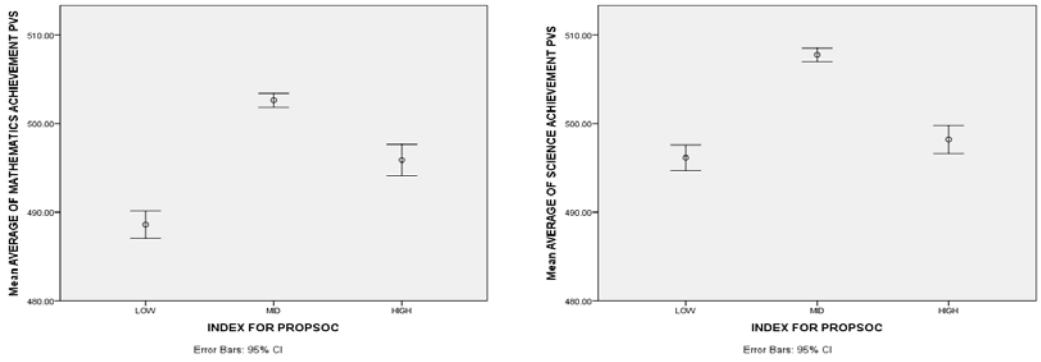


Figure 4.1-3: Mean achievement scores by index of proportion social



Appendix 4.2. Classroom experiences (G4 mathematics)

Explaining answers

Over fifty per cent of respondents have limited opportunities to 'explain their answers' in class, reporting either never or only in some lessons. In terms of association with mathematics achievement, there is evidence in Table 4.2-1 to suggest that higher scores are associated with increased opportunities to explain their thinking, although when this demand goes beyond half the lessons, with an expectation that one explains answers 'every or almost every lesson', there is a drop in mathematics achievement score.

Table 4.2-1: Average of mathematics achievement by explain my answer

			MATHEMATICS	
MAT\HOW OFTEN\I EXPLAIN MY ANSWERS\REVERSE	N	Percent	Mean	S.D.
NEVER	6,600	12.5	474.3	78.2
SOME LESSONS	22,200	41.9	501.3	71.6
ABOUT HALF THE LESSONS	13,100	24.8	507.2	72.2
EVERY OR ALMOST EVERY LESSON	11,000	20.8	482.2	77.9
Total	52,900	100.0	495.4	74.9

Memorising

The findings on experiences that involve an element of rote learning, where students memorise how to work out problems or are expected to memorise formulae and procedures are presented in Table 4.2-2. A small proportion of students, around 6% of the cohort, report 'never' memorising formulae or procedures; respondents in this group have the lowest mathematics achievement scores. There is a broadly even distribution of response across the other categories that are all associated with comparable mathematics achievement scores, with no one group reflecting a particularly high score as witnessed with the other variables under consideration in this section.

Table 4.2-2: Average of mathematics achievement by I memorise formulae and procedures

			MATHEMATICS	
MAT\HOW OFTEN\I WORK IN GROUPS\REVERSE	N	Percent	Mean	S.D.
NEVER	3,300	6.2	464.9	84.2
SOME LESSONS	15,800	29.9	494.6	71.9
ABOUT HALF THE LESSONS	16,000	30.2	499.7	72.0
EVERY OR ALMOST EVERY LESSON	17,900	33.7	498.4	76.1
Total	53,000	100.0	495.6	74.6

Practical Activities and Presentation of Data

Nearly 70 per cent of respondents report doing measuring tasks in some of their lessons; fewer than twenty per cent of students report a higher uptake, with the highest frequency of 'every or almost every lesson' accounting for less than ten per cent of the cohort. The highest take up level of this experience is associated with the lowest achievement scores. The highest achievement scores are associated with the dominant category of 'some lessons'.

Table 4.2-3: Average of mathematics achievement by I measure things in class

MAT\HOW OFTEN\I MEASURE THINGS IN CLASS\REVERSE	N	Percent	MATHEMATICS Mean	S.D.
NEVER	7,500	14.1	483.8	76.9
SOME LESSONS	35,800	67.6	507.6	71.1
ABOUT HALF THE LESSONS	5,200	9.9	475.8	69.6
EVERY OR ALMOST EVERY LESSON	4,400	8.4	439.6	73.3
Total	53,000	100.0	495.4	74.8

A similar picture is portrayed for the variable that measures producing tables, charts and graphs. The highest achievement scores are associated with 'some lessons', the dominant response for the experience, but the lowest achievement scores are associated with those who 'never' have such an opportunity. Over a third of respondents report a high frequency of experience with data representation tasks (making tables, charts and graphs) in half of their lessons or more. Table 4.2-4 shows the associated achievement scores are reduced for higher frequency of experience, but not to the levels as witnessed with measure tasks in Table 4.2-3.

Table 4.2-4: Average of mathematics achievement by I make tables charts & graphs

MAT\HOW OFTEN\MAKE TABLES CHARTS GRAPHS\REVERSE	N	Percent	MATHEMATICS Mean	S.D.
NEVER	4,200	7.9	451.2	82.2
SOME LESSONS	29,300	55.3	507.9	70.3
ABOUT HALF THE LESSONS	12,500	23.5	500.1	70.9
EVERY OR ALMOST EVERY LESSON	7,000	13.3	461.5	74.3
Total	53,000	100.0	495.4	74.7

Working by myself; Working with others

The next two student-level variables relate to classroom organisation, illustrating whether and how group or individual activity is associated with achievement scores. Working on problems in class, pursuing individual tasks on one's own, is reported by over half the respondents as occurring every or almost every lesson. This frequency of experience is associated with high mathematics achievement scores as documented in Table 4.2-5. A broadly linear association between frequency of experience and achievement scores is observed across 'some' to 'every or almost every' lesson in Figure 4.2-1, with a large drop-off in achievement score witnessed within the small group who 'never' experience working problems on their own.

Table 4.2-5: Average of mathematics achievement by I work problems on own

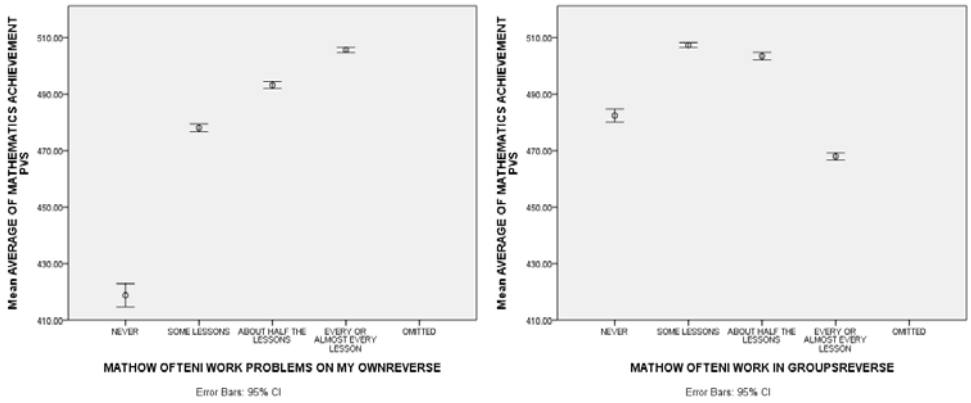
MAT\HOW OFTEN\I WORK PROBLEMS ON MY OWN\REVERSE	N	Percent	MATHEMATICS Mean	S.D.
NEVER	1,300	2.4	418.7	74.5
SOME LESSONS	10,500	19.8	478.0	72.2
ABOUT HALF THE LESSONS	11,800	22.3	493.1	69.1
EVERY OR ALMOST EVERY LESSON	29,500	55.6	505.6	74.9
Total	53,000	100.0	495.3	74.8

Nearly half of the respondents report doing group work on 'some lessons', a feature that is associated with high mathematics achievement scores. Just under a quarter reports doing group work on 'every or almost every lesson', but this frequency of experience is associated with lower achievement scores than those observed across the middle categories that extends to 'about half the lessons'. The small proportion of respondents who 'never' do group work appear to be associated with low achievement scores, all as illustrated in Table 4.2-6 and Figure 4.2-1.

Table 4.2-6: Average of mathematics achievement by I work in small groups

MAT\HOW OFTEN\I WORK IN GROUPS\REVERSE	N	Percent	MATHEMATICS Mean	S.D.
NEVER	3,500	6.6	482.4	70.5
SOME LESSONS	25,500	47.9	507.2	69.9
ABOUT HALF THE LESSONS	11,300	21.2	503.4	74.7
EVERY OR ALMOST EVERY LESSON	12,900	24.3	467.9	78.1
Total	53,200	100.0	495.2	74.9

Figure 4.2-1: MEAN MATH ACHIEVEMENT (95% CI) BY PROBLEMS ON MY



Daily life

The last variable in this analysis of G4 mathematics experiences concerns the prevalence of undertaking work related to daily life. This is based on a teacher-level variable where the class teacher indicates how often they ask students to relate what they are learning to their daily lives. Just over one fifth of students have a teacher who provides for this classroom experience on a very regular basis ('every or nearly every lesson'). Nearly half the cohort can expect this experience in 'some lessons' but there is no obvious association between frequency of experience and mathematics achievement scores.

Table 4.2-7: Average of mathematics achievement by ask to relate to daily lives

			MATHEMATICS	
MAT\HOW OFTEN\I ASK TO RELATE TO DAILY LIVES\REVERSE	N	Percent	Mean	S.D.
SOME LESSONS	22,300	46.2	497.8	76.6
ABOUT HALF THE LESSONS	16,000	33.1	487.6	73.6
EVERY OR ALMOST EVERY LESSON	10,000	20.7	499.3	72.1
Total	48,400	100.0	494.7	74.9

Appendix 4.3. Classroom experiences (G4 science)

Explaining

Around one sixth of the cohort report 'never' having opportunities to 'explain something studied in science'. There is evidence in Table 4.3-1 to suggest students in that position, who never explain their studies, tend to be associated with low science achievement scores. Where there are opportunities for those types of interactions, even if only occasionally, the experience is associated with higher student achievement scores in science; higher that is until this demand goes as far 'as at least once a week', at which point there is a drop in the associated science achievement score, much as witnessed with mathematics achievement scores.

Table 4.3-1: Average of science achievement by explain something studied

			SCIENCE	
SCI\HOW OFTEN\GIVE EXPLAN STUDYING SCI\REVERSE	N	Percent	Mean	S.D.
NEVER	8,900	16.7	482.9	73.6
A FEW TIMES A YEAR	13,400	25.4	508.4	64.8
ONCE OR TWICE A MONTH	18,100	34.1	509.1	70.0
AT LEAST ONCE A WEEK	12,600	23.8	494.3	73.4
Total	53,000	100.0	501.0	70.9

Memorising facts

Relative to findings for mathematics, nearly three times as many students report 'never' having to memorise science facts, a feature that is associated with low achievement scores. High science achievement scores are associated with higher responses on this item with minimal diminution even when memorising activity is expected at least once a week.

Table 4.3-2: Average of science achievement by memorise science facts

			SCIENCE	
SCI\HOW OFTEN\MEMORIZE SCIENCE FACTS\REVERSE	N	Percent	Mean	S.D.
NEVER	8,900	17.0	474.4	68.3
A FEW TIMES A YEAR	11,600	22.1	497.2	67.4
ONCE OR TWICE A MONTH	15,900	30.3	513.1	68.5
AT LEAST ONCE A WEEK	16,000	30.6	507.9	73.0
Total	52,400	100.0	501.4	70.9

Experiments and investigations

The findings are reported in Table 4.3-3 through Table 4.3-5. One clear message that comes across in all three scenarios is the association between students 'never' having opportunities to engage with science experiments or investigations and low science achievement scores. The proportion of students reporting the 'never' category is also noteworthy: with nearly one third of respondents never having responsibility for 'planning' a science experiment or investigation; over a fifth report never 'doing' experiments, and a slightly smaller proportion (17%) claim never to 'watch' their teacher do a science experiment. Comparable achievement scores are associated with the middle categories of 'a few times a year' and 'once or twice a month' for each of the three experiences.

Table 4.3-3: Average of science achievement by WATCH science experiment or investigation

SCI\HOW OFTEN\WATCH TEACHER DO SCI EXP\REVERSE	N	Percent	SCIENCE Mean	S.D.
NEVER	9,000	17.0	482.3	72.8
A FEW TIMES A YEAR	13,800	26.0	517.0	67.2
ONCE OR TWICE A MONTH	14,800	27.9	514.0	66.7
AT LEAST ONCE A WEEK	15,500	29.1	484.3	71.3
Total	53,200	100.0	500.7	71.1

Table 4.3-4: Average of science achievement by PLAN science experiment or investigation

SCI\HOW OFTEN\ I PLAN SCI EXP OR INVESTIGATION\REVERSE	N	Percent	SCIENCE Mean	S.D.
NEVER	16,700	31.7	490.1	67.9
A FEW TIMES A YEAR	16,200	30.8	513.3	70.0
ONCE OR TWICE A MONTH	12,600	23.9	506.2	70.8
AT LEAST ONCE A WEEK	7,200	13.6	486.4	76.9
Total	52,600	100.0	500.6	71.4

Table 4.3-5: Average of science achievement by DO science experiment or investigation

SCI\HOW OFTEN\ I DO SCI EXP OR INVESTIGATION\REVERSE	N	Percent	SCIENCE Mean	S.D.
NEVER	11,500	21.8	479.8	72.1
A FEW TIMES A YEAR	16,900	32.1	513.1	63.8
ONCE OR TWICE A MONTH	14,400	27.4	510.9	70.5
AT LEAST ONCE A WEEK	9,900	18.7	489.1	74.9
Total	52,700	100.0	500.7	71.1

Increasing levels of practical work for students, whether watching their teacher, planning the experiment themselves, or doing the experiment, are associated with decreasing science achievement scores with a pronounced dip when those experiences occur 'at least once a week'.

However, cross tabulations indicate otherwise. Focusing on the 'never' category, the response that is associated with low achievement scores, the partial breakdown of planning science experiments is presented in Table 4.3-6. 'Planning' is selected for this analysis because it is the experience that had least exposure, with nearly a third of students never engaging with that aspect of practical work in science. Although there is a high overlap of responses in the 'never' category, over 25 per cent of respondents who never do science experiments report some level of input on planning science experiments or investigations. Similarly, around 40 per cent of those who never watch the teacher do experiments indicate opportunities to plan.

Table 4.3-6: Analysis of PLAN science experiment with 'never' DO or WATCH

		PLAN SCIENCE EXPERIMENT OR INVESTIGATION				
		AT LEAST ONCE A WEEK	ONCE OR TWICE A MONTH	A FEW TIMES A YEAR	NEVER	Total
DO SCI EXP OR INVEST	NEVER	500 (4%)	500 (4%)	2,000 (18%)	8,400 (74%)	11,400
WATCH TEACHER DO SCI EXPERIMENT	NEVER	500 (6%)	800 (9%)	2,200 (25%)	5,400 (61%)	8,900

If watching a teacher do a science experiment is regarded as a minimum entitlement for science practical activities, the data presented in Table 4.3-7 shows that this is not met for over 40 per cent of respondents. Those students never do science experiments themselves, and do not see experiments demonstrated by their teacher. Over two thirds of those who never plan experiments or investigations report watching their teacher do experiments, leaving a third neither planning nor watching.

Table 4.3-7: Analysis of WATCH teacher do science experiment with 'never' DO or PLAN

		WATCH TEACHER DO SCIENCE EXPERIMENT				
		AT LEAST ONCE A WEEK	ONCE OR TWICE A MONTH	A FEW TIMES A YEAR	NEVER	Total
DO SCI EXP OR INVEST	NEVER	2,000 (18%)	1,700 (15%)	2,900 (26%)	4,800 (42%)	11,300
PLAN SCI EXP OR INVEST	NEVER	3,100 (19%)	3,600 (22%)	4,600 (28%)	5,400 (33%)	16,600

To complete the picture of analysis for this cluster of experiences surrounding practical work in science, an analysis of 'doing science experiments' is presented in Table 4.3-8. Some fifty five per cent of respondents who never watch their teacher do experiments do not themselves do science experiments or investigations. A similar proportion of those who never plan experiments do not personally do science experiments.

Table 4.3-8: Analysis of DO science experiment with 'never' WATCH or PLAN

		DO SCIENCE EXPERIMENT OR INVESTIGATION				
		AT LEAST ONCE A WEEK	ONCE OR TWICE A MONTH	A FEW TIMES A YEAR	NEVER	Total
WATCH TEACHER DO SCI EXPERIMENT	NEVER	600 (7%)	900 (10%)	2,500 (28%)	4,800 (55%)	8,800
PLAN SCI EXP OR INVEST	NEVER	1,600 (10%)	2,200 (13%)	4,200 (26%)	8,400 (51%)	16,400

These high proportions of students who do not actively engage with science experiments or investigations (neither watch, plan nor do), even at a low level of a few times a year, appear to be disadvantaged by nature of the association between those experiences and science achievement scores reported above.

Working by myself; Working with others

The next pair of variables concerns organisational aspects of learning and the effect of working on science problems either alone or collaboratively. The distribution of responses and associated science achievement scores are presented in Table 4.3-9 and Table 4.3-10. The central positions for each variable of 'a few times a year' and up to 'once or twice a month' appear to be associated with higher science achievement scores, increasing as the frequency of experience increases. Fewer than 10 per cent of respondents never work on their own when studying science problems, with the bulk of students regularly experiencing this type of classroom activity or organisation.

Experience of working in groups is not so prevalent for the G4 students with around 10 per cent 'never' working with others on experiments or investigations and a further quarter of the cohort only doing so a few times a year. Working in groups on experiments is associated with high achievement scores; highest when experienced once or twice a month but the association stands even if experienced only on a few occasions in the year.

Table 4.3-9: Average of science achievement by I work science problems on own

			SCIENCE	
SCI\HOW OFTEN\ I WORK SCI PROBLEMS ON OWN\REVERSE	N	Percent	Mean	S.D.
NEVER	5,200	9.7	469.2	74.0
A FEW TIMES A YEAR	9,400	17.7	505.1	66.7
ONCE OR TWICE A MONTH	16,800	31.5	509.8	68.5
AT LEAST ONCE A WEEK	21,900	41.1	499.4	72.2
Total	53,300	100.0	500.8	71.2

Table 4.3-10: Average of science achievement by I work in group experiment or investigation

			SCIENCE	
SCI\HOW OFTEN\ I WORK IN GROUP EXPERIMENTS OR INVES\REVERSE	N	Percent	Mean	S.D.
NEVER	5,900	11.1	476.4	76.1
A FEW TIMES A YEAR	13,400	25.2	509.6	66.3
ONCE OR TWICE A MONTH	17,600	33.1	513.9	68.4
AT LEAST ONCE A WEEK	16,300	30.7	487.6	71.8
Total	53,200	100.0	500.6	71.2

Daily life

The last variable in this analysis of experiences, concerns work related to daily life. This variable is based on a teacher-level contribution as outlined above for mathematics.

Table 4.3-11: Average of science achievement by ask to relate to daily life

			SCIENCE	
SCI\HOW OFTEN\ ASK TO RELATE TO DAILY LIFE\REVERSE	N	Percent	Mean	S.D.
SOME LESSONS	15,600	34.2	501.5	73.7
ABOUT HALF THE LESSONS	16,300	35.6	499.7	71.8
EVERY OR ALMOST EVERY LESSON	13,800	30.2	502.0	69.2
Total	45,600	100.0	501.0	71.7

Just under one third of students have a teacher who relates science lessons to daily life on a very regular basis ('every or nearly every lesson'). In comparison to experiences with their mathematics teachers, a higher proportion of students can expect this type of experience in 'about half the lessons' or more frequently, with a fairly even spread of response across the three categories reported. There is no obvious association between frequency of experience and science achievement scores, other than noting that the achievement scores for all three reporting categories are relatively consistent and close to the grand mean; there is no obvious dip in association in the tails of the distribution as witnessed with other variables and that all teachers certainly ask students to relate what they are learning in science to their daily lives (N=0 for 'never').

Technology

Around one third of students report 'never' using a calculator in their mathematics lessons, a feature that is associated with lower achievement scores than those associated with using a calculator in 'some lessons'. A small proportion of students use calculators more regularly than that, with less than 10 per cent using a calculator in about half their lessons or more frequently; higher frequency of calculator use is associated with ever decreasing achievement scores.

Table 4.3-12: Average of mathematics achievement by use of calculator

MAT\HOW OFTEN\I USE A CALCULATOR IN CLASS\REVERSE			MATHEMATICS	
	N	Percent	Mean	S.D.
NEVER	18,100	33.9	485.0	75.1
SOME LESSONS	30,600	57.4	506.9	71.7
ABOUT HALF THE LESSONS	3,500	6.5	474.5	71.9
EVERY OR ALMOST EVERY LESSON	1,200	2.3	412.2	71.0
Total	53,300	100.0	495.2	74.9

A high proportion of students report 'never' using a computer in mathematics classes – nearly 50 per cent of the cohort – with just over a further third using a computer in some of their lessons. There is no obvious association between frequency of experience and mathematics achievement scores, other than noting that the achievement scores dip in line with very frequent use of computers, i.e. when this extends to every or almost every mathematics lesson.

Table 4.3-13: Average of G4 mathematics achievement by use of computer in class

MAT\HOW OFTEN\I USE A COMPUTER IN CLASS\REVERSE			MATHEMATICS	
	N	Percent	Mean	S.D.
NEVER	25,000	46.8	500.3	72.3
SOME LESSONS	19,800	37.2	500.8	72.6
ABOUT HALF THE LESSONS	4,900	9.1	492.9	73.5
EVERY OR ALMOST EVERY LESSON	3,700	6.9	434.7	78.0
Total	53,400	100.0	495.3	74.8

Extending this analysis to take account of student's use of computers in support of schoolwork, in and out of school, gives the distribution of use and associated achievement scores presented in Table 4.3-14. Over one third of the cohort report using a computer at least once a week or more (every day) for schoolwork, a situation that is associated with low achievement scores. The highest achievement scores are associated with fairly infrequent use, a slightly different pattern of distribution and association with achievement scores than witnessed above where the analysis was restricted to in-school-use, but similar to the extent that high use of computers does not appear to offer any advantages in terms of achievement score and those who 'never' use computers for their mathematics schoolwork, in or out of school, have a higher than average achievement score.

Table 4.3-14: How often I use computer for school work in and out of school (mathematics)

MAT\HW OFTEN USE COMPUTER FOR SCHWORK\MATH			MATHEMATICS	
	N	Percent	Mean	S.D.
NEVER	15,000	28.7	500.4	75.2
A FEW TIMES A YEAR	8,500	16.2	517.3	70.2
ONCE OR TWICE A MONTH	10,300	19.7	506.9	70.2
AT LEAST ONCE A WEEK	14,600	27.9	484.6	70.5
EVERY DAY	3,900	7.5	450.7	77.1
Total	52,300	100.0	496.3	74.3

Data on use of ICT in Science is restricted to use of a computer; no data was collected on frequency of calculator use in science lessons. More than half the cohort never uses a computer in science lessons with just over a fifth using a computer in some lessons. As with the mathematics data there is no obvious association between frequency of experience and science achievement scores, other than noting that the achievement scores dip with very frequent use of computers in science lessons. The 10 per cent of students who use a computer 'every or almost every lesson' have low science achievement scores.

Table 4.3-15: Average of science achievement by use of computer in class

SCI\HOW OFTEN\I USE A COMPUTER IN SCIENCE LESSONS\REVERSE			SCIENCE	
	N	Percent	Mean	S.D.
NEVER	27,600	51.8	500.9	70.2
SOME LESSONS	11,800	22.2	515.0	69.8
ABOUT HALF THE LESSONS	8,600	16.1	505.4	67.9
EVERY OR ALMOST EVERY LESSON	5,300	9.9	462.1	69.6
Total	53,300	100.0	500.9	71.1

Student's use of computers in support of science schoolwork, in and out of school, is presented in Table 4.3-16. The highest achievement scores are associated with fairly infrequent use, a very similar pattern to that witnessed in support of mathematics schoolwork, with over a third 'never' using computers for school work in or out of school. This is in contrast to the earlier data on computer availability at home, where the majority of students (95%) reported possessing a computer at home; the data in Table 4.3-14 and Table 4.3-16 tell us that home computers are not used extensively to directly support school work whether internet connected or not.

Table 4.3-16: How often I use computer for school work in and out of school (science)

SCI\HW OFTEN USE COMPUTER FOR SCHOOL WORK\SCI			SCIENCE	
	N	Percent	Mean	S.D.
NEVER	19,100	37.2	498.5	69.0
A FEW TIMES A YEAR	11,400	22.1	520.3	65.4
ONCE OR TWICE A MONTH	10,900	21.2	512.9	70.5
AT LEAST ONCE A WEEK	8,000	15.5	483.3	67.1
EVERY DAY	2,000	3.9	444.1	73.8
Total	51,400	100.0	501.9	70.5

Appendix 4.4. School culture and ethos (G4)

Measures of school culture and ethos are derived from the two clusters of variables that relate first to students' perception of peers' attitudes to study and teachers' expectation of students, and second how welcoming and 'safe' they find the school environment.

The climate of learning is reported through two variables that seek response on the extent to which: 'I think that children at my school try to do their best'; and 'I think that teachers at my school want children to do their best'. Responses for both items range from 'disagree a lot' through to 'agree a lot'.

Students try their best

Respondents clearly agree with the statement, with fewer than 10 per cent disagreeing and over 60 per cent who 'agree a lot'. Students' views on their peers putting in their best effort do not appear to be strongly associated with their mathematics or science achievement scores.

Table 4.4-1: Average of mathematics and science achievement by students try to do their best

GEN\AGREE\STUDENTS TRY THEIR BEST			MATHEMATICS		SCIENCE	
	N	Percent	Mean	S.D.	Mean	S.D.
DISAGREE A LOT	900	1.7	451.2	89.2	475.2	86.1
DISAGREE A LITTLE	3,400	6.4	514.7	80.6	517.0	75.5
AGREE A LITTLE	16,200	30.4	513.0	70.5	518.1	65.7
AGREE A LOT	32,900	61.5	485.7	73.6	491.3	70.7
Total	53,400	100.0	495.3	74.8	500.8	71.1

Teachers want students to do their best

The students' perceptions of their teacher's expectations are strongly supportive of the view that teachers want the best for children in their care, with over 90 per cent agreeing 'a lot'. In Table 4.4-2 there appears to be an association between achievement scores and perceived strength of feeling on a positive and supportive culture, where teachers show that they want the best for their learners; low achievements scores are associated with disagreement, although only a very small proportion of respondents do not feel their teacher wants their children to do their best.

Table 4.4-2: Average of mathematics and science achievement by teachers want students to do their best

GEN\AGREE\TEACHERS WANT STUDENTS TO DO THEIR BEST			MATHEMATICS		SCIENCE	
	N	Percent	Mean	S.D.	Mean	S.D.
DISAGREE A LOT	500	1.0	436.1	94.6	438.5	99.4
DISAGREE A LITTLE	600	1.1	461.4	89.9	472.0	78.4
AGREE A LITTLE	2,800	5.3	500.7	78.6	509.4	74.5
AGREE A LOT	49,400	92.7	496.1	73.7	501.4	70.1
Total	53,300	100.0	495.4	74.8	500.9	71.1

The climate of learning is taken to be the combination of these two variables, but given the skewed nature of the separate distributions a dichotomous derived variable is generated to reflect agreement with both variables ('a little' or 'a lot') or disagreement in one. The distribution and association with achievement scores is presented in Table 4.4-3, where data show a stronger association with mathematics achievement scores than that reported for science achievement.

Table 4.4-3: Average of mathematics and science achievement by climate (student perception)

STUDENT CLIMATE \TRY THEIR BEST & TEACHERS WANT BEST			MATHEMATICS		SCIENCE	
	N	Percent	Mean	S.D.	Mean	S.D.
DISAGREE ON ONE	5,000	9.3	492.8	88.7	500.1	83.0
AGREE ON BOTH	48,300	90.7	495.7	73.1	501.1	69.7
Total	53,300	100.0	495.4	74.7	501.0	71.1

The second measure of school ethos reflects how 'safe' students feel at school, taking into account their responses to five features of school life. An IEA derived variable produced an index of High, Medium and Low perception of being safe, based on the students' responses to the following set of questions:

In school, did any of these things happen during the last month?

1. Something of mine was stolen (AS4GSTOL);
2. I was hit or hurt by other student(s) (e.g., shoving, hitting, kicking) (AS4GHURT);
3. I was made to do things that I didn't want to do by other students (AS4GMADE);
4. I was made fun of or called names (AS4GMFUN);
5. I was left out of activities by other students (AS4GLEFT).

A high perception of safety was assigned on the basis of responding 'no' to all five statements; a low perception of safety assigned if there were three or more responses of 'yes'; with all other combinations falling into the medium category for perception of safety. [The

index was coded as missing if there were 2 or more source questions with invalid data.] Table 4.4-4 shows an association between feeling safe in school and achievement, with lower achievement scores when students report a low perception of safety.

Table 4.4-4: Average of mathematics and science achievement by student perception of safety index

IDX STD PRCPTN BEING SAFE IN SCHOOL (SPBSS)			MATHEMATICS		SCIENCE	
	N	Percent	Mean	S.D.	Mean	S.D.
HIGH	21,400	40.1	501.0	70.8	505.4	67.3
MEDIUM	20,800	38.9	499.8	74.2	506.2	69.1
LOW	11,200	21.0	475.8	80.2	481.9	78.3
Total	53,400	100.0	495.2	74.8	500.8	71.1

Two fifths of the cohort reports no concerns on safety, answering 'no' to all five statements. A closer analysis of achievement scores by the number of concerns (from zero to all five issues as presented in Table 4.4-5; ASDSAFE4 variable based on at least 4 responses from the 5) does not add a great deal to the derived variable above. Only a small proportion of students report four or more concerns (~10%) and the cut-point of three or more can be seen to signify a step change in average achievement scores in mathematics and science. For those reasons and to retain a parsimonious model, the index version of safety (ASDGPBSS) will be used in subsequent analyses.

Table 4.4-5: Average of mathematics and science achievement by students' perception of being safe

STUD PERCEPTION OF BEING SAFE (STOLEN, HURT, FORCED, FUN, IGNORED)			MATHEMATICS		SCIENCE	
	N	Percent	Mean	S.D.	Mean	S.D.
0 CONCERNS	21,400	40.1	501.0	70.8	505.4	67.3
1 OF 5 CONCERNS	11,900	22.3	503.4	72.4	509.8	67.2
2 OF 5 CONCERNS	8,900	16.6	494.8	76.2	501.3	71.4
3 OF 5 CONCERNS	6,100	11.4	481.4	77.8	488.8	74.4
4 OF 5 CONCERNS	3,400	6.4	471.0	75.8	477.0	75.9
ALL 5 CONCERNS	1,800	3.3	465.8	94.2	467.4	92.5
Total	53,400	100.0	495.2	74.8	500.8	71.1

Appendix 4.5. Background data on G8 students

Appendix 4.5.1. Biographical data

An even gender balance is reported in G8 with girls making up 50.9% of the weighted sample, but the associated mathematics and science achievement scores highlight a gender imbalance with boys scoring significantly higher than girls in the same stage. The spread of achievement scores is smaller for girls than boys in both subject disciplines, with girls having more clustered responses as shown in Table 4.5-1. This finding concurs with that research evidenced discussed in Chapter 5.1.1 (Powney, 1996 and SSLN, 2012).

Table 4.5-1: AVERAGE achievement by GENDER (BSDGSEX)

		Mathematics		Science	
GENDER	N	Mean	Std. Deviation	Mean	Std. Deviation
BOY	28,900	489.4	78.1	499.1	79.5
GIRL	30,000	486.9	75.4	493.7	75.0
Total	58,800	488.1	76.8	496.3	77.3

Figures have been rounded so columns may not equal total

A range of 'home' factors are now presented, taking account of each variable's distribution and association with achievement. A first cluster of variables takes account of parental background and language spoken at home, and a second cluster serves as a proxy for social and economic status of the household. For the first aspect, a new variable is computed to combine data on students' place of birth and where they spent their early formative years. This variable is an extension of that presented within G4 analyses with an additional category to cater for those students who came to UK when older than 10 years. The three categories provide data on whether or not a student was born in UK, and if not, whether they came to UK before school age (≤5 years), during their primary education phase (aged 5 to 10 years), or when older than 10 years. The standard naming convention for G8 variables (as deployed by IEA) is to have the prefix 'B' – in this particular case the derived variable is BSDBFORM, a G8 student-variable derived from other registered variables, with the same naming convention as described earlier.

Table 4.5-2: Formative years in UK (BSDBFORM)

BORN OR LIVED IN UK FOR FORMATIVE YEARS (≤5YRS)		Frequency	Percent	Cumulative Percent
Valid	EARLY FORMATIVE YEARS SPENT IN UK	55,900	96.8	96.8
	ARRIVED IN UK AGED 5 TO 10 YEARS	700	1.2	98.0
	ARRIVED IN UK OLDER THAN 10 YEARS	1200	2.0	100.0
	Total	57,800	100.0	

Around 3% of respondents arrived in UK older than five years; the mean mathematics and science achievement scores are significantly lower for those students. The data in Table 4.5-3 show the feature of 'not having spent early formative years in the native country' has a strong association with mathematics achievement score; science achievement is similarly associated with this student variable, albeit with a higher overall achievement in science. A clear pattern of lower achievement scores appears to be associated with students who arrived in the UK in their later years, with those arriving after age 10 being associated with the lowest achievement scores in mathematics and science.

Table 4.5-3: AVERAGE of achievement by Early Formative Years (BSDBFORM)

		Mathematics		Science	
BORN OR LIVED IN UK FOR FORMATIVE YEARS (<5YRS)	N	Mean	Std. Deviation	Mean	Std. Deviation
FORMATIVE YEARS SPENT IN UK	55,900	490.5	75.5	498.7	75.9
ARRIVED IN UK AGED 5 TO 10 YEARS	700	452.3	90.3	459.6	93.9
ARRIVED IN UK OLDER THAN 10 YEARS	1,200	436.1	94.3	439.2	92.5
Total	57,800	488.9	76.6	497.1	77.1

Figures have been rounded so columns may not equal total

The home-based variable concerning spoken language used at home (BS4OLANG) is reported with associated mean achievement scores in G8 mathematics and science. Fewer than 10% of students do not 'always' speaking English at home, and the mean achievement scores for students in those categories is considerably lower than that reported by others. Unlike the data from the G4 students, there is no enhancement within the 'almost always' category; Table 4.5-4 shows a pattern of declining achievement associated with reducing levels of English spoken at home.

Table 4.5-4: AVERAGE of achievement by Own Language at Home (BS4OLANG)

			Mathematics		Science	
OWN LANGUAGE AT HOME	N	Percent	Mean	s.d.	Mean	s.d.
ALWAYS	53,000	90.3	490.1	75.9	498.8	76.1
ALMOST ALWAYS	3,300	5.6	485.6	80.3	493.8	78.1
SOMETIMES	1,700	2.9	463.0	81.4	464.4	87.6
NEVER	700	1.3	416.8	61.4	411.4	60.0
Total	58,700	100.0	488.2	76.7	496.4	77.2

The next home-based variables, or environmental measures to be considered are the 'number of books in the home', a proxy for parental interest and potential to engage student through modelling study techniques and personal reading at home, and resources available to students in support of their studies. Following a principal component analysis the latter was broken down into 'study tools' (calculator, dictionary, and encyclopaedia) and ICT resources (computer and internet) that were available to students in their home (Components 1and 2 in Table 4.5-5). As with the G4 data, the PCA analysis resulted in components being loaded on a single variable for possession of 'mobile phone', and 'own bedroom' as demonstrated in Table 4.5-5: Rotated Component Matrixa (G8 Resources). Possessing a 'study desk' was dropped from the analysis because it was loaded on more than one component, following the same line of argument as set out in Chapter 5.1.3.

Table 4.5-5: Rotated Component Matrix[a] (G8 Resources)

	Component			
	1	2	3	4
GEN\HOME POSSESS\CALCULATOR	0.746			
GEN\HOME POSSESS\COMPUTER		0.864		
GEN\HOME POSSESS\DICTIONARY	0.808			
GEN\HOME POSSESS\INTERNET CONNECTION		0.840		
GEN\HOME POSSESS\OWN BEDROOM				0.975
GEN\HOME POSSESS\MOBILE PHONE			0.970	
GEN\HOME POSSESS\ENCYCLOPAEDIA	0.680			

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
Factor loadings of less than 0.4 are not shown
a. Rotation converged in 4 iterations.

Students were asked to estimate the number of books in their home with the reporting framework coded into five categories as before: 0 to 10 books; 11 to 25 books, 26 to 100 books; 101 to 200 books, and more than 200 books. Just under one half of G8 students report having 25 or fewer books at home, a feature that is almost double that reported at G4. These categories are associated with low mean achievement scores as illustrated in Table 4.5-6. An increasing benefit from having more books at home appears to be highlighted in the association with science achievement scores, where the mean scores are higher than those reported in mathematics and the spread of scores is lower with smaller standard deviations reported within the corresponding categories. On this occasion, with the older students, there is not the same degree of levelling off or plateauing of achievement scores witnessed at G4 when students reported over 100 books at home (two or more bookcases); increased number of books is associated with higher achievement scores in both disciplines.

Table 4.5-6: Average of achievement scores by number of books (BS4BOOKS)

			Mathematics		Science	
NUM BOOKS AT HOME	N	Percent	Mean	s.d.	Mean	s.d.
0 TO 10 BOOKS	12,900	22.1	439.3	66.4	436.1	66.3
11 TO 25 BOOKS	14,000	24.0	468.7	64.1	477.5	60.0
26 TO 100 BOOKS	14,900	25.5	498.8	68.9	507.6	64.3
101 TO 200 BOOKS	8,200	13.9	526.9	67.8	538.4	64.3
OVER 200 BOOKS	8,500	14.6	539.8	75.6	560.6	70.7
Total	58,600	100.0	488.3	76.7	496.6	77.2

Around ten per cent of students had none or one of the study tools, whereas nearly two thirds had all three of calculator, dictionary and encyclopaedia at their disposal for home study.

Table 4.5-7: Average of achievement by Study Tools (calculator, dictionary, and encyclopaedia)

			Mathematics		Science	
POSSESS\STUDY TOOLS CALC, DICT, ENCYCL (BSDTOOLS)	N	Percent	Mean	s.d.	Mean	s.d.
NO STUDY TOOLS	2,000	3.4	429.9	66.7	423.4	63.9
ONE OF CALC, DICT, ENCYCLOPAEDIA	4,000	6.9	448.0	65.7	444.0	69.1
TWO OF CALC, DICT, ENCYCLOPAEDIA	14,000	23.9	472.6	70.7	477.3	69.0
ALL THREE OF CALC, DICT, ENCYCLOPAEDIA	38,300	65.8	501.9	76.1	513.5	74.9
Total	58,300	100.0	488.7	76.6	497.0	77.1

There is a fairly linear association between the number of study tools and mean achievement scores, with higher mean scores associated with having all three study tools at home (Table 4.5-7).

A significant majority of the G8 students (97.5%) report having a home computer. Within that group only 6% do not have an internet connection. This measure is taken as an indicator of home support along the lines of 'number of books'. The empirical data in Table 4.5-8 shows possession of a computer *with* internet connection is positively associated with mean mathematics and science achievement. As with the other home resources, the effect of increased support in the form of ICT has a stronger association with science achievement scores.

Table 4.5-8: Average of achievement by ICT (COMPUTER/INTERNET)

			Mathematics		Science	
POSSESS\ICT (COMPUTER+/-INT)	N	Percent	Mean	s.d.	Mean	s.d.
NO HOME COMPUTER	1,400	2.5	436.1	66.8	439.0	67.6
COMPUTER WITHOUT INTERNET	3,600	6.1	448.8	75.6	458.4	79.7
HOME COMPUTER AND INTERNET	53,600	91.5	492.6	75.5	500.9	75.7
Total	58,600	100.0	488.6	76.5	496.8	77.0

The remaining components within the home resource cluster, reflect on possession of 'own bedroom' and possession of a 'mobile phone'.

A substantial majority of the students report having their own bedroom (85%), a feature that has the potential to be associated with having a study space at home. This variable on 'possessing own bedroom' could also serve as an indicator of SES. The empirical data in Table 4.5-9 shows an association between having one's own bedroom and both mathematics and science achievement scores. A stronger association appears to be present within the science achievement data.

Table 4.5-9: Average of achievement by Own Bedroom (BS4GTH06)

			Mathematics		Science	
POSSESS OWN BEDROOM	N	Percent	Mean	s.d.	Mean	s.d.
YES	49,700	84.6	490.9	76.9	500.6	76.8
NO	9,000	15.4	473.9	73.9	474.4	75.4
Total	58,700	100.0	488.3	76.7	496.5	77.2

Possession of a mobile phone within the G4 student cohort was high, with over 85% reporting ownership. This rises to 97% of G8 students, a feature that appears to be associated with a small but significant *positive* effect on student achievement (F=9.7; p=0.002); Table 4.5-10 illustrates that association, reversing the direction of effect reported within G4 data.

Table 4.5-10: Average of mathematics achievement by Mobile Phone (BS4GTH07)

			Mathematics		Science	
POSSESS MOBILE PHONE	N	Percent	Mean	s.d.	Mean	s.d.
YES	57,000	97.0	488.4	76.4	496.7	76.7
NO	1,800	3.0	482.6	85.1	489.4	91.0
Total	53,600	100.0	488.2	76.7	496.5	77.2

Appendix 4.5.2. Out-of-school interests and experiences (G8)

A similar analysis of the out-of-school interests and experiences for students in G8 produced a different pattern of relationship. Each student reported how much time was spent on each of the nine activities – the eight discussed in Chapter 5.1.4 (watching TV or videos; playing computer games; playing or talking with friends; doing jobs at home; playing sports; reading books for enjoyment; using the internet; and doing homework) plus an additional variable on level of work undertaken in a paid job. Responses to the nine items were subjected to a principal component analysis using ones as prior communality estimates. The principal axis method was used to extract the components, and this was followed by a varimax (orthogonal) rotation as discussed earlier.

In interpreting the rotated factor pattern, an item was taken to load on a given component if the factor loading was 0.4 or greater for that component, and was less than 0.4 for the other components. The resultant components are slightly different to those reported for G4 students, but still reflect on out-of-school activities as primarily: of an 'individual' nature (IND); as a reflection on 'personal development' (PDEV); and as a measure of 'social' engagement with other students (SOC); plus two standalone components that loaded on a single variable, i.e. Play Sport (PLSP) and Paid Work (WKPJ)

The 'social' dimension with this age group identified two variables as contributors for the component: play and talk with friends; and using the internet. This represents a shift in loading from that reported for G4 students, where 'using the internet' was aligned with solitary

activities and grouped within the 'individual' component. For the older students the data shows a higher loading on the 'social' component, with likely links to social media on the internet (such as Facebook); I note in Table 4.5-11 that there remains a small loading on the 'individual' component (0.28) but that this loading is not included in subsequent analyses; dropped from consideration on ground of loading being less than 0.4.

Table 4.5-11: Rotated Component Matrix[a] (G8 Outside School)

Things you do outside school	Component		
	1 (IND)	2 (PDEV)	3 (SOC)
GEN\SPEND TIME\WATCH TV OR VIDEOS	0.80		
GEN\SPEND TIME\PLAY COMPUTER GAMES	0.85		
GEN\SPEND TIME\PLAY TALK WITH FRIENDS			0.86
GEN\SPEND TIME\DO JOBS AT HOME		0.66	
GEN\SPEND TIME\READ BOOK FOR ENJOYMENT		0.73	
GEN\SPEND TIME\USE INTERNET	(0.28)		0.71
GEN\SPEND TIME\DO HOMEWORK		0.76	

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
Factor loadings of less than 0.4 are not shown
a. Rotation converged in 4 iterations.

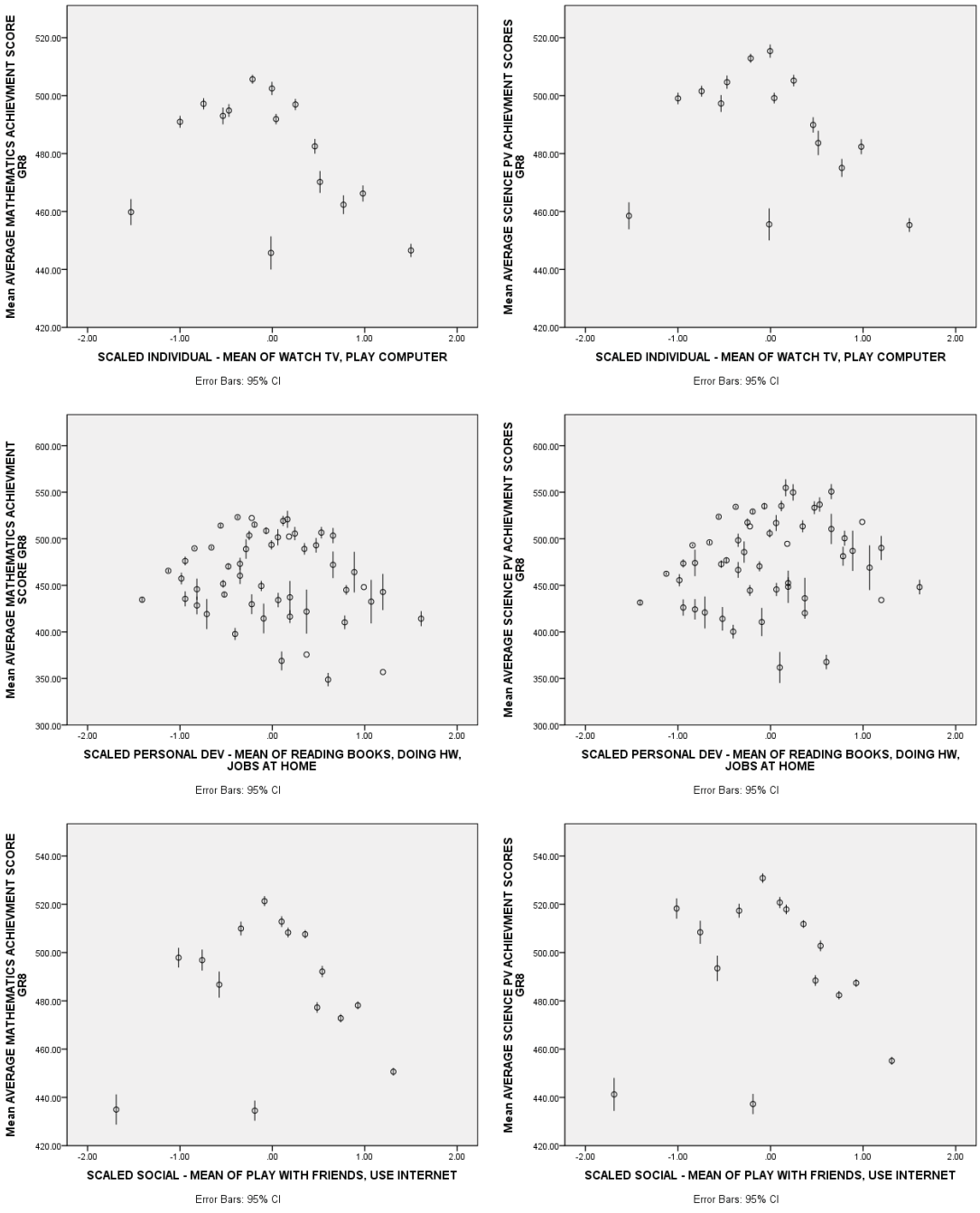
These three components that reflect 'individual', 'personal development' and 'social' aspects of out-of-school activities are re-scaled in line with arguments presented earlier (Snell, 1964). The re-scaling factors for the score values contributing to each of components are as shown in Table 4.5-12.

Table 4.5-12: Re-scale factors for INDIVIDUAL (IND), PERSONAL DEVELOPMENT (PDEV) and SOCIAL (SOC) components

	j	1	2	3	4	5
IND	Normalised	-1.53	-0.47	0.04	0.46	1.50
PDEV	Normalised	-1.41	-0.56	-0.01	0.37	1.61
SOC	Normalised	-1.69	-0.34	0.17	0.54	1.31

The resultant associations with mathematics and science achievement scores are shown in Figure 4.5-1. These charts support a curvilinear association with achievement scores for both distributions, highlighting similar patterns across disciplines albeit on different scales with the science scores generally higher than the equivalent mathematics achievement scores plotted against the scaled responses on out-of-school activities. When these variables are entered into the multilevel models the best fitting polynomial will be selected to maximise potential association with the dependent variable.

Figure 4.5-1: Distribution of 'out-of-school' components against achievement scores



The two stand-alone components for 'playing sport' and having a 'paid job' were reported on the basis of time spent each school day, ranging from no time through to 4 or more hours. Just under a third of respondents indicate less than an hour or no time playing sport, with almost a fifth engaged in sporting activities for 4 or more hours each day. Table 4.5-13 shows an association with high mean achievement scores in both mathematics and science where students report playing sport for an hour, to less than four hours a day; anything longer than that

is associated with considerably lower achievement scores. There is a distinct difference in uptake across genders (percentage of male /female participation rates are included in Table 4.5-13) but there are comparable numbers of boys and girls in the response categories that appear to offer the greatest association with enhanced achievement scores i.e. 1 to 2 hours and the neighbouring category of more than 2 but less than 4 hours. The empirical data supports the argument for an active schools agenda, encouraging participation in a broad definition of sport such as that offered by Council of Europe (2001):

> "Sport means all forms of physical activity which, through casual and organised participation, aim at expressing or improving physical fitness and mental well-being, forming social relationships or obtaining results in competition at all levels."

Coulter (2005) notes there is no definitive evidence of a positive, causal relationship between physical activity/sport and academic achievement and although there are some studies where correlations have been found (Thomas et al , 1994; Etnier et al, 1997), the explanation for the nature and direction of association remains speculative. Coulter concludes that although the evidence is inconclusive on improved academic performance, there are indications that it does not have a negative effect and sporting activity does offer physical and emotional benefits. The empirical data here is less supportive of that final claim if high levels of participation are witnessed, where a negative association is documented when playing sport for 4 or more hours per day.

Table 4.5-13: Average of achievement by 'playing sport'

GEN\SPEND TIME\I PLAY SPORTS					Mathematics		Science	
	N	Percent	M(%)	F(%)	Mean	s.d.	Mean	s.d.
NO TIME	7,900	13.7	7.7	19.5	472.0	80.5	483.6	83.0
LESS THAN I HOUR	10,900	19.0	13.0	24.8	487.5	75.5	494.7	77.3
1 TO 2 HOURS	16,600	28.8	25.9	31.6	503.2	75.5	510.3	75.9
MORE THAN 2 BUT LESS THAN 4 HOURS	10,700	18.6	23.1	14.3	501.4	73.1	511.4	73.1
4 OR MORE HOURS	11,400	19.8	30.3	9.8	470.1	72.2	477.7	70.5
Total	57,400	100.0	100.0	100.0	489.0	76.4	497.4	76.9

Over three-quarters of the students do not work in a paid capacity. Of those that do, there appears to be a negative association with achievement scores in both mathematics and science as the level of engagement in paid work increases. Table 4.5-14: Average of achievement by 'paid job', shows the small percentage of respondents within each category beyond 'no time', and the associated average achievement scores. From the empirical data on

working in a paid job, this type of activity appears to impart a greater effect on mathematics achievement scores over science scores.

Table 4.5-14: Average of achievement by 'paid job'

			Mathematics		Science	
GEN\SPEND TIME\WORK PAID JOB	N	Percent	Mean	s.d.	Mean	s.d.
NO TIME	43,600	76.3	495.0	74.5	503.0	75.2
LESS THAN I HOUR	4,000	7.1	486.0	77.9	490.6	78.9
1 TO 2 HOURS	4,200	7.3	480.8	74.8	483.6	76.7
MORE THAN 2 BUT LESS THAN 4 HOURS	2,300	4.1	474.0	77.2	482.8	78.6
4 OR MORE HOURS	3,000	5.2	439.6	76.1	462.6	78.1
Total	57,100	100.0	489.6	76.1	497.8	76.7

Appendix 4.6. Classroom experiences (G8 mathematics)

Respondents were asked to indicate the frequency of exposure to the experience or style of learning, using a 4-point scale that runs from 'every or almost every lesson' through 'about half the lessons' and 'some lessons', down to 'never'. Responses on the scale were 'reversed' to present higher scores in line with the claimed benefit in learning and understanding. An analysis of frequency of exposure is presented alongside the mean mathematics achievement score for each variable; analyses of experiences in science are presented separately because the range of experiences is not replicated across disciplines.

Taking each experience in turn, the findings are:

Explaining answers

Over fifty per cent of G8 respondents explain their answers in class 'every or almost every lesson'. This represents a big change from their G4 counterparts where only a fifth reported that level of opportunity. The evidence in Table 4.6-1 points to increased achievement scores as the level of explaining increases, where explaining for half the lessons or more is associated with higher achievement scores in mathematics. Fewer than five per cent of respondents claim to 'never' explain their answers, a feature that is associated with the lowest mean achievement scores.

Table 4.6-1: Average of mathematics achievement by explain my answer

MAT\HOW OFTEN\I EXPLAIN MY ANSWERS\REVERSE	N	Percent	MATHEMATICS Mean	S.D.
NEVER	2,700	4.7	472.7	74.5
SOME LESSONS	11,800	20.4	486.7	74.6
ABOUT HALF THE LESSONS	12,300	21.3	493.3	75.9
EVERY OR ALMOST EVERY LESSON	31,100	53.7	490.8	77.0
Total	57,900	100.0	489.0	76.3

Memorising

The findings on students' level of memorising formulae and procedures are presented in Table 4.6-2. Around double the proportion of G4 students report 'never' memorising in their G8 class, a situation that is associated with the lowest achievement scores. There is evidence to support an association between memorising and mathematics achievement, with scores increasing in tandem with level of opportunity or expectation to memorise formulae and processes. As opportunities to memorise in class increase to 'every or almost every lesson' the effect and association with achievement scores is reduced.

Table 4.6-2: Average of mathematics achievement by memorise formulae and procedures

MAT\HOW OFTEN\I MEMORISE FORMULAE ETC. \REVERSE	N	Percent	MATHEMATICS Mean	S.D.
NEVER	8,000	13.7	450.0	68.5
SOME LESSONS	24,400	42.1	493.1	73.3
ABOUT HALF THE LESSONS	14,500	25.0	501.8	74.3
EVERY OR ALMOST EVERY LESSON	11,000	19.1	493.1	81.2
Total	57,900	100.0	489.4	76.2

Solving complex problems

The association between levels of ownership of procedures for solving complex problems with achievement scores shows a marked reduction in achievement as the frequency of opportunity increases. This empirical finding is at odds with claims on the benefits of autonomy and ownership of strategies used within problem solving, where students who frequently decide on their own procedures for solving complex problems are associated with lower achievement scores. The highest achievement scores are associated with the largest response category of 'some lessons', reflecting the benefits of having some opportunity to take ownership of the process but the highest take up of this experience is associated with the lowest achievement scores; unusually lower than the 'never' category.

Table 4.6-3: Average of mathematics achievement by solving complex problems

MAT\HOW OFTEN\SOLVING COMPLEX PROBLEMS	N	Percent	MATHEMATICS Mean	S.D.
NEVER	8,900	15.4	483.5	73.5
SOME LESSONS	25,900	45.0	495.6	71.7
ABOUT HALF THE LESSONS	14,500	25.2	489.1	81.2
EVERY OR ALMOST EVERY LESSON	8,300	14.4	473.8	83.3
Total	57,600	100.0	489.0	76.6

Interpreting Data

Nearly two thirds of respondents report interpreting data in some of their lessons, a level of engagement that is associated with the highest achievement scores. Any increase in time devoted to interpreting data in graphs, charts and tables is associated with lower achievement scores. Clearly an imbalance in curricular opportunities will be restrictive on overall achievement in mathematics. The highest take up level of this experience is associated with the lowest achievement scores, again noted as lower than the 'never' category.

Table 4.6-4: Average of mathematics achievement by interpreting data

MAT\HOW OFTEN\I MEASURE THINGS IN CLASS\REVERSE	N	Percent	MATHEMATICS Mean	S.D.
NEVER	2,400	4.2	449.8	64.9
SOME LESSONS	37,700	64.8	501.3	72.3
ABOUT HALF THE LESSONS	12,200	21.0	479.8	78.9
EVERY OR ALMOST EVERY LESSON	5,800	10.0	443.7	76.3
Total	58,100	100.0	488.8	76.4

Working with others

The two student-level variables related to classroom organisation illustrate whether and how group or individual activity is associated with achievement scores. Working on problems on own, pursuing individual tasks in class, is reported as occurring 'every or almost every lesson' by over one third of respondents. Much as reported by students in G4, this frequency of experience is associated with high mathematics achievement scores as documented in Table 4.6-5. A broadly linear association between frequency of experience and achievement scores is observed across the four response categories.

Table 4.6-5: Average of mathematics achievement by work problems on own

MAT\HOW OFTEN\I WORK PROBLEMS ON MY OWN\REVERSE	N	Percent	MATHEMATICS Mean	S.D.
NEVER	2,600	4.4	450.2	67.1
SOME LESSONS	15,200	26.2	471.0	72.9
ABOUT HALF THE LESSONS	20,200	35.0	490.7	72.7
EVERY OR ALMOST EVERY LESSON	19,900	34.4	506.2	78.5
Total	57,900	100.0	489.1	76.3

Over a third of the students claim 'never' to work together in small groups, with just over two-fifths of the respondents doing group work on 'some lessons'. The latter category is associated with high mathematics achievement scores as shown in Table 4.6-6. It is worth noting that higher levels of participation in group work are associated with lower achievement scores, lower even than the scores associated with 'never' category. Fewer than ten per cent report doing group work on 'every or almost every lesson', but this frequency of experience is associated with the lowest achievement scores across all response categories.

Table 4.6-6: Average of mathematics achievement by work in small groups

MAT\HOW OFTEN\I WORK IN GROUPS\REVERSE	N	Percent	MATHEMATICS Mean	S.D.
NEVER	20,700	35.7	490.0	71.5
SOME LESSONS	24,400	42.2	497.4	76.6
ABOUT HALF THE LESSONS	8,200	14.1	487.9	78.9
EVERY OR ALMOST EVERY LESSON	4,700	8.1	442.8	76.9
Total	53,200	100.0	495.2	74.9

Daily life

Just under one fifth of students never experience relating what they learn in mathematics to their daily life, but this does not seem to adversely affect their learning in that the 'never' category is associated with a higher than average achievement score. Those who experience links to daily life in 'some lessons' are associated with similar mathematics achievement scores, whilst any higher levels of engagement are associated with lower achievement scores as presented in Table 4.6-7. Although the absolute difference in achievement scores is not as extreme as witnessed with some of the other variables, the differences between groups are still statistically significant ($F_{3, 57912} = 418.1$, $p < 0.001$).

Table 4.6-7: Average of mathematics achievement by learning related to daily life

MAT\HOW OFTEN\I ASK TO RELATE TO DAILY LIVES\REVERSE	N	Percent	MATHEMATICS Mean	S.D.
NEVER	10,000	17.3	498.6	70.5
SOME LESSONS	21,900	37.8	496.2	75.6
ABOUT HALF THE LESSONS	14,400	24.9	488.5	76.1
EVERY OR ALMOST EVERY LESSON	11,600	20.0	468.0	79.5
Total	57,900	100.0	489.1	76.5

Reviewing homework

All three response categories that report experiencing review of homework are associated with higher than average achievement scores in mathematics. Over one third of the respondents report this experience as a regular feature of their classroom activity, happening on every or almost every lesson. Less than fifteen per cent of the G8 students 'never' review their homework in class; this shortfall on formative assessment is associated with low achievement scores.

Table 4.6-8: Average of mathematics achievement by review homework

MAT\HOW OFTEN\WE REVIEW OUR HOMEWORK_REVERSE	N	Percent	MATHEMATICS Mean	S.D.
NEVER	7,900	13.7	462.5	78.1
SOME LESSONS	16,300	28.2	492.8	71.6
ABOUT HALF THE LESSONS	13,500	23.4	498.5	73.8
EVERY OR ALMOST EVERY LESSON	20,000	34.7	490.8	78.7
Total	57,700	100.0	489.2	76.3

Quiz or test

Another source of formative assessment comes through the use of a quiz or test in class and, much as above, the category of students who 'never' experience this type of activity is associated with low mathematics achievement scores. Just over three-quarters of the students report a quiz or test as a feature of their learning experience in 'some lessons', and this appears to be associated with the highest achievement scores. When the use of a quiz or test increases to be a very regular feature of class work, happening every or almost every lesson, the benefits appear to be lost in that this category is associated with the lowest achievement scores; lower than 'never', although the small proportion reporting in that category places a restriction on the accuracy of the estimate. This regular use of a quiz or test relates to classroom situations where assessment begins to take over at the expense of learning, with the instrument or experience potentially no longer viewed as a primarily formative exercise.

Table 4.6-9: Average of mathematics achievement by quiz or test

MAT\HOW OFTEN\HAVE A QUIZ OR TEST_REVERSE	N	Percent	MATHEMATICS Mean	S.D.
NEVER	1,100	2.0	457.6	80.8
SOME LESSONS	43,700	75.2	496.8	74.1
ABOUT HALF THE LESSONS	7,900	13.6	483.7	75.6
EVERY OR ALMOST EVERY LESSON	5,400	9.3	440.7	73.9
Total	58,100	100.0	489.0	76.4

Lecture-style presentation

A high proportion of G8 students report listening to the teacher giving a lecture-style presentation on 'every or nearly every lesson'. This experience is associated with a high achievement score, whereas the lowest mathematics achievement score is associated with the 'never' category. Again there is a highly significant variance across categories where decreasing achievement scores are associated with the lower frequencies of a lecture-style of presentation ($F_{3, 58203} = 1113.5$, $p < 0.001$).

Table 4.6-10: Average of mathematics achievement by lecture-style presentation

MAT\HOW OFTEN\LISTEN TEACHER LECTURE_REVERSE	N	Percent	MATHEMATICS Mean	S.D.
NEVER	1,700	3.0	423.5	75.8
SOME LESSONS	8,200	14.1	460.4	66.7
ABOUT HALF THE LESSONS	10,800	18.5	484.3	73.0
EVERY OR ALMOST EVERY LESSON	37,500	64.4	499.5	76.2
Total	58,200	100.0	488.9	76.4

Technology

A very small proportion of students 'never' use a calculator in mathematics classes in comparison to over half the respondents using a calculator for 'about half the lessons' or more frequently; the latter categories being associated with high achievement scores as illustrated in Table 4.6-11.

Table 4.6-11: Average of mathematics achievement by use of calculator in G8 class

MAT\HOW OFTEN\I USE A CALCULATOR IN CLASS\REVERSE	N	Percent	MATHEMATICS Mean	S.D.
NEVER	1,900	3.3	429.9	74.5
SOME LESSONS	25,500	43.7	476.1	71.3
ABOUT HALF THE LESSONS	18,800	32.2	508.0	73.0
EVERY OR ALMOST EVERY LESSON	12,100	20.8	495.3	82.5
Total	58,200	100.0	488.9	76.4

In contrast to calculator usage, a large proportion of students 'never' use a computer in their Grade 8 mathematics class. This is higher than reported by G4 students (Table 4.3-13), but the main difference in the G8 group is that the highest achievement scores are associated with never using a computer in class, rather than with the 'never' or 'some' categories that were associated with high achievement in the G4 data.

Table 4.6-12: Average of mathematics achievement by use of computer in G8 class

MAT\HOW OFTEN\I USE A COMPUTER IN CLASS\REVERSE			MATHEMATICS	
	N	Percent	Mean	S.D.
NEVER	39,400	67.7	497.4	70.6
SOME LESSONS	14,300	24.5	483.9	81.8
ABOUT HALF THE LESSONS	2,600	4.4	434.5	78.0
EVERY OR ALMOST EVERY LESSON	1,900	3.3	415.6	81.5
Total	58,100	100.0	488.6	76.6

Extending this analysis to take account of student's use of computers in support of schoolwork, in and out of school, gives the distribution of use and associated achievement scores presented in Table 4.6-13 (Mathematics). Generally most students report 'never' using a computer in support of their mathematics studies (over 50%); but highest achievement scores are associated with low levels of computer use, up to once or twice a month.

Table 4.6-13: Average of mathematics achievement by use of computer for mathematics sch work

MAT\HW OFTEN USE COMPTR FOR SCHWORK\MATH			MATHEMATICS	
	N	Percent	Mean	S.D.
NEVER	31,000	53.9	488.9	73.7
A FEW TIMES A YEAR	10,900	19.0	511.6	73.3
ONCE OR TWICE A MONTH	9,400	16.4	495.0	73.4
AT LEAST ONCE A WEEK	5,000	8.6	457.0	75.4
EVERY DAY	1,200	2.1	399.3	77.3
Total	57,600	100.0	489.6	76.2

Appendix 4.7. Classroom experiences (G8 science)

Explaining

Only a small proportion of G8 students 'never' have opportunities to give explanations about what they are studying. There is evidence in Table 4.7-1 to suggest students in that position tend to be associated with lower science achievement scores. Where there are opportunities for those interactions, the experience is associated with higher student achievement scores in science; highest when occurring every or almost every lesson. There is broadly an even spread of response across the three other categories with about a third of the cohort reporting 'about half the lessons' or 'every or almost every lesson'.

Table 4.7-1: Average of science achievement by explain something studied

			SCIENCE	
SCI\HOW OFTEN\GIVE EXPLANATION OF STUDYING SCI\REVERSE	N	Percent	Mean	S.D.
NEVER	3.800	6.6	466.6	83.4
SOME LESSONS	15,900	27.6	486.9	74.5
ABOUT HALF THE LESSONS	19,300	33.5	501.0	75.0
EVERY OR ALMOST EVERY LESSON	18,600	32.3	508.7	76.0
Total	57,500	100.0	497.4	76.7

Memorising facts

The distribution of responses to this experience of memorising facts and principles is broadly evenly spread, with a third of the cohort reporting 'some lessons' or 'about half the lessons'. The latter category is associated with the highest achievement score in science (Table 4.7-2). Those students reporting memorising as a regular activity, on every or almost every lesson, show a reduction in achievement score that mirrors reported changes in mathematics achievement for the same variable.

For G8 students there appears to be shift in emphasis that supports opportunities to explain over memorising, with higher achievement scores associated with regular exposure to explaining what is studied. This is a different picture to that reported earlier (G4 science) where memorising appeared to take precedence over explaining in its association with high achievement scores.

Table 4.7-2: Average of science achievement by memorise science facts

SCI\HOW OFTEN\MEMORIZE SCIENCE FACTS & PRINCIPLES\REVERSE	N	Percent	SCIENCE Mean	S.D.
NEVER	4,200	7.3	466.8	81.5
SOME LESSONS	19,400	33.6	494.7	77.0
ABOUT HALF THE LESSONS	19,500	33.8	506.1	73.5
EVERY OR ALMOST EVERY LESSON	14,600	25.3	498.8	76.9
Total	57,600	100.0	497.6	76.8

Experiments and investigations

The next three experiences relate to science experiments or investigations, drawing a distinction between opportunities to 'watch' the teacher demonstrate the activity and students actively 'planning' or 'conducting' the experiment or investigation themselves. The findings are reported in Table 4.7-3 through Table 4.7-5. A significant change from data reported on G4 student concerns the level of response in the 'never' category. For the older students only small proportions of respondents fall into that category of never engaging with experiments. In all three scenarios (watch, plan and conduct) the students who 'never' have opportunities to engage with science experiments or investigations are associated with low science achievement scores. The proportion of students reporting engagement is evenly spread across the three response categories, with roughly one third of respondents affirming engagement. For each of the three variables the highest achievement scores are associated with the middle category of 'about half the lessons'. The same strength and direction of association is noted for the higher category of 'every or almost every lesson' within conducting experiments, whereas the empirical data for higher levels of 'watching' and 'planning' are associated with slightly lower achievement scores.

Table 4.7-3: Average of science achievement by watch teacher DEMONSTRATE

SCI\HOW OFTEN\WATCH TEACHER DEMONSTRATE SCI EXP\REVERSE	N	Percent	SCIENCE Mean	S.D.
NEVER	1,300	2.2	444.7	89.8
SOME LESSONS	15,100	25.9	489.1	77.1
ABOUT HALF THE LESSONS	20,400	35.1	507.5	75.3
EVERY OR ALMOST EVERY LESSON	21,400	36.8	495.6	75.7
Total	58,200	100.0	497.0	77.0

Table 4.7-4: Average of science achievement by PLAN science experiment or investigation

SCI\HOW OFTEN\ I PLAN SCI EXP OR INVESTIGATION\REVERSE	N	Percent	SCIENCE Mean	S.D.
NEVER	3,900	6.8	470.1	82.2
SOME LESSONS	18,400	31.9	498.0	75.3
ABOUT HALF THE LESSONS	19,300	33.5	506.0	78.2
EVERY OR ALMOST EVERY LESSON	16,100	27.8	492.5	74.1
Total	57,700	100.0	497.3	77.0

Table 4.7-5: Average of science achievement by CONDUCT science experiment or investigation

SCI\HOW OFTEN\ I DO SCI EXP OR INVESTIGATION\REVERSE	N	Percent	SCIENCE Mean	S.D.
NEVER	1,300	2.3	436.2	75.9
SOME LESSONS	12,900	22.5	482.8	76.9
ABOUT HALF THE LESSONS	18,800	32.7	504.3	75.6
EVERY OR ALMOST EVERY LESSON	24,400	42.5	502.9	75.3
Total	57,400	100.0	497.3	76.7

Increasing levels of practical work for students, whether watching their teacher, planning the experiment themselves, or conducting experiments, are associated with increasing science achievement scores where the experience occurs in about half the lessons or more. Much as reported for G4 students on this aspect of classroom experience, there is a distinct disadvantage for students who never benefit from such opportunities to engage with scientific practical work; fortunately this only affects small proportions of the cohort but the differences across categories remain statistically significant with 'conducting experiments' providing the greatest explanation of variance:

Watch teacher demonstrate ($F_{3, 58236} = 388.2$, $p < 0.001$).

Plan a science experiment ($F_{3, 57683} = 270.9$, $p < 0.001$).

Conduct experiment or investigation ($F_{3, 57423} = 539.1$, $p < 0.001$).

Working by myself; Working with others

The next pair of variables concerns organisational aspects of learning and the effect of working on science problems either alone or collaboratively. The distribution of responses and associated science achievement scores are presented in Table 4.7-6 and Table 4.7-7. The proportion of students reporting 'never' has reduced from the levels documented for G4 science, with a particularly noticeable reduction from 11.1% to 2.0% for working in small groups. Only

8 per cent of respondents never work on their own when studying science problems, with the bulk of students experiencing this type of classroom activity or organisation in at least some lessons. The highest achievement scores are associated with working individually on problems in about half the lessons.

Table 4.7-6: Average of science achievement by work science problems on own

			SCIENCE	
SCI\HOW OFTEN\ I WORK SCI PROBLEMS ON OWN\REVERSE	N	Percent	Mean	S.D.
NEVER	4,700	8.1	472.5	82.3
SOME LESSONS	21,100	36.4	492.3	74.1
ABOUT HALF THE LESSONS	20,300	35.0	507.2	74.9
EVERY OR ALMOST EVERY LESSON	11,900	20.5	498.0	80.0
Total	58,000	100.0	497.1	76.9

Experience of working in groups is a dominant classroom activity with over half the cohort reporting this as happening in every or almost every lesson. A small proportion of G8 students never work in groups; this situation is associated with very low achievement scores. Table 4.7-7 shows that increased opportunities to work collaboratively are associated with ever improving achievement scores. Over 80% of respondents are associated with a higher than average achievement score in science which is quite a contrast to that reported on this variable and its association with mathematics achievement where: (a) much smaller proportions (<25%) indicated engagement at those levels; and (b) the associated achievement scores were lower than in other categories, including the 'never' option that was reported by over a third of the students.

Table 4.7-7: Average of science achievement by work in small groups

			SCIENCE	
SCI\HOW OFTEN\ I WORK IN GROUP EXPERIMENTS OR INVES\REVERSE	N	Percent	Mean	S.D.
NEVER	1,100	2.0	426.8	79.9
SOME LESSONS	10,300	17.8	481.4	74.0
ABOUT HALF THE LESSONS	16,300	28.1	500.7	79.1
EVERY OR ALMOST EVERY LESSON	30,300	52.2	503.1	74.8
Total	58,000	100.0	497.0	77.1

Daily life

This variable concerns how often students relate what they are learning in science to their daily lives. Fewer than 15% 'never' do this in their lessons, a position that is associated with low achievement scores.

Table 4.7-8: Average of science achievement by relate to daily life

SCI\HOW OFTEN\ ASK TO RELATE TO DAILY LIFE\REVERSE	N	Percent	SCIENCE Mean	S.D.
NEVER	8,200	14.2	485.9	75.8
SOME LESSONS	21,400	37.2	499.3	74.8
ABOUT HALF THE LESSONS	17,400	30.3	501.9	76.4
EVERY OR ALMOST EVERY LESSON	10,600	18.4	495.9	82.1
Total	57,600	100.0	497.6	77.0

There is no obvious association between frequency of experience and science achievement scores, other than noting that the achievement scores for all three reporting categories beyond 'never' are relatively consistent and close to the grand mean; there is no obvious dip in association in the tails of the distribution as witnessed with some other variables and the majority of students appear to benefit from linking what they are learning in science to their daily lives; a significant association even if only happening in some lessons. ($F_{3, 57549} = 86.8$, $p < 0.001$).

Reviewing homework

Over one fifth of G8 students never experience review of science homework, with a further third only reporting this activity has happening in 'some lessons'. However, as the data in Table 4.7-9 highlights, those categories are associated with high science achievement scores. Nearly one quarter of the respondents report this experience as a regular feature of their classroom activity, happening on every or almost every lesson, but this category is associated with the lowest mean achievement score, which calls into question the value of such experience, or the manner in which it is administered in the science classroom. By comparison this type of formative assessment was deemed to be worthwhile in mathematics classes where there was a clear step change from 'never' to the other categories of response.

Table 4.7-9: Average of science achievement by review homework

SCI\HOW OFTEN\WE REVIEW OUR HOMEWORK_REVERSE	N	Percent	SCIENCE Mean	S.D.
NEVER	12,500	21.8	500.9	78.1
SOME LESSONS	18,800	32.6	500.5	78.7
ABOUT HALF THE LESSONS	13,200	23.0	498.7	71.7
EVERY OR ALMOST EVERY LESSON	12,900	22.5	488.7	77.2
Total	57,500	100.0	497.5	76.8

Quiz or test

The category of student who 'never' experiences this type of activity is associated with low mathematics achievement scores. Over 70% of the G8 students experience this type of formative assessment in at least some lessons, an experience that is associated with high science achievement scores.

Table 4.7-10: Average of science achievement by quiz or test

SCI\HOW OFTEN\HAVE A QUIZ OR TEST_REVERSE	N	Percent	SCIENCE Mean	S.D.
NEVER	1,800	3.0	449.8	77.6
SOME LESSONS	41,000	70.6	507.3	74.5
ABOUT HALF THE LESSONS	9,900	17.0	488.4	74.8
EVERY OR ALMOST EVERY LESSON	5,400	9.3	452.4	74.9
Total	58,000	100.0	497.3	76.9

As with the data on mathematics achievement, when the use of a quiz or test increases to be a very regular feature of class work, happening every or almost every lesson, the benefits appear to be lost in that this response is associated with the lowest achievement scores; even lower than the 'never' category.

Lecture-style presentation

A high proportion of G8 students report listening to the teacher giving a lecture-style presentation on 'every or nearly every lesson' in science. This experience is associated with a high achievement score, whereas the lowest mathematics achievement score is associated with the 'never' category. Again there is a highly significant variance across categories with decreasing achievement scores associated with lower exposure to a lecture-style of presentation ($F_{3, 57732} = 1133.3$, $p < 0.001$).

Table 4.7-11: Average of science achievement by lecture-style presentation

SCI\HOW OFTEN\LISTEN TEACHER LECTURE_REVERSE	N	Percent	SCIENCE Mean	S.D.
NEVER	1,700	2.9	439.2	77.5
SOME LESSONS	8,200	14.1	469.8	71.0
ABOUT HALF THE LESSONS	12,700	22.0	487.2	72.7
EVERY OR ALMOST EVERY LESSON	35,200	61.0	510.2	76.1
Total	57,800	100.0	497.4	76.8

Technology

Analysis of student's use of computers in support of schoolwork, in and out of school, gives the distribution of use and associated achievement scores presented in Table 4.7-12. In comparison to data presented on mathematics, a similar pattern of usage is reported across disciplines with just over ten per cent of the students using computers more than 'once or twice a month', an experience that is associated with low achievement scores. Generally fewer students report 'never' using a computer in support of their science studies compared to mathematics studies (35% in science with over 50% mathematics); but highest achievement scores in both disciplines are associated with low levels of computer use, up to once or twice a month.

Table 4.7-12: Average of science achievement by use of computer for science sch work

MAT\HW OFTEN USE COMPTR FOR SCHWORK\SCIENCE	N	Percent	SCIENCE Mean	S.D.
NEVER	19,900	35.0	478.9	76.6
A FEW TIMES A YEAR	15,100	26.5	517.1	71.1
ONCE OR TWICE A MONTH	15,600	27.3	510.3	73.3
AT LEAST ONCE A WEEK	5,300	9.4	486.3	80.2
EVERY DAY	1,100	1.9	451.1	83.8
Total	56,900	100.0	497.8	76.9

Appendix 4.8. School culture and ethos (G8)

The climate of learning is reported through responses to: 'I think that children at my school try to do their best'; and 'I think that teachers at my school want children to do their best'; responses ranging from 'disagree a lot' through to 'agree a lot'.

Students try their best

Three-fifths of respondents agree with the statement, with fewer than 10 per cent disagreeing a lot. Students' views on their peers putting in their best effort do not appear to be strongly associated with their mathematics or science achievement scores in Table 4.8-1.

Table 4.8-1: Average of mathematics and science achievement by students try to do their best

GEN\AGREE\STUDENTS TRY THEIR BEST			MATHEMATICS		SCIENCE	
	N	Percent	Mean	S.D.	Mean	S.D.
DISAGREE A LOT	5,600	9.7	473.7	75.9	485.4	80.6
DISAGREE A LITTLE	17,700	30.4	495.6	71.0	508.3	71.6
AGREE A LITTLE	25,300	43.4	495.8	75.2	497.8	75.6
AGREE A LOT	9,600	16.5	466.0	84.1	480.1	84.5
Total	58,200	100.0	488.7	76.5	496.9	77.1

Teachers want students to do their best

In Table 4.8-2 there appears to be an association between achievement scores and perceived strength of feeling on a positive and supportive culture, where teachers show that they want the best for their learners; low achievements scores are associated with disagreement (a little, or a lot), although only a small proportion of respondents (~10%) do not feel their teacher wants their children to do their best.

Table 4.8-2: Average of mathematics and science achievement by teachers want students to do their best

GEN\AGREE\TEACHERS WANT STUDENTS TO DO THEIR BEST			MATHEMATICS		SCIENCE	
	N	Percent	Mean	S.D.	Mean	S.D.
DISAGREE A LOT	2,400	4.1	446.5	74.2	438.3	76.0
DISAGREE A LITTLE	3,700	6.4	477.3	76.2	478.0	77.8
AGREE A LITTLE	15,700	26.9	495.0	73.4	502.8	74.2
AGREE A LOT	36,500	62.6	490.0	77.1	500.3	76.0
Total	53,300	100.0	488.8	76.5	497.0	77.1

The second measure of school ethos reflects how 'safe' students feel at school, an IEA derived index-variable of High, Medium and Low perception of being safe as described in 0.

The comparable variables within the G8 data have the 'B' prefix.

A high perception of safety was assigned on the basis of responding 'no' to all five concerns; a low perception of safety assigned if there were three or more concerns over 'safety'.

Table 4.8-3: Average of mathematics and science achievement by student perception of safety index

IDX STD PRCPTN BEING SAFE IN SCHOOL (SPBSS)			MATHEMATICS		SCIENCE	
	N	Percent	Mean	S.D.	Mean	S.D.
HIGH	34,700	59.5	490.5	74.9	495.3	75.4
MEDIUM	18,800	32.3	489.2	77.7	501.7	78.7
LOW	4,700	8.1	474.2	82.6	490.7	82.2
Total	58,200	100.0	488.8	76.6	497.0	77.1

Sixty per cent of the cohort reports no concerns on safety. Where there are concerns, medium or low, those categories are associated with lower achievement scores, more notably within the mathematics discipline.

Appendix 5. Additional Tables from Primary Analysis

Appendix 5.1. Model of science achievement (G4) with ALL control and predictor variables

Table 5.1-1: Modelling all predictor and control variables (G4 Science - Controls)

ASSSCIPV	ASSCI_ALL_COMBINED_MODEL_IGLS				
Response	Coef.	S.E.	Sig	effect size	+/- error
FORMATIVE YEARS IN COUNTRY(UK)					
ARRIVED IN UK AGED 5 OR OVER	-39.8	4.8	**	-0.66	-0.08
OWN LANGUAGE					
ALMOST ALWAYS	10.1	2.8	**	0.17	0.05
SOMETIMES	-19.8	4.1	**	-0.33	-0.07
NEVER	-23.5	7.4	**	-0.39	-0.12
STUDY TOOLS					
ONE OF CALC, DICT, ENCYCLOPAEDIA	8.3	7.7	-	0.14	0.13
TWO OF CALC, DICT, ENCYCLOPAEDIA	14.0	7.2	-	0.23	0.12
ALL THREE OF CALC, DICT, ENCYCLOPAEDIA	18.6	7.2	**	0.31	0.12
HOME ICT					
COMPUTER WITHOUT INTERNET	2.4	5.3	-	0.04	0.09
COMPUTER WITH INTERNET	26.0	4.4	**	0.43	0.07
BOOKS AT HOME					
NONE OR VERY FEW (0 TO 10 BOOKS)	-28.1	3.7	**	-0.47	-0.06
ONE SHELF (11 TO 25 BOOKS)	-14.8	2.7	**	-0.25	-0.05
TWO BOOKCASES (101 TO 200 BOOKS)	16.5	2.6	**	0.27	0.04
THREE OR MORE BOOKCASES (OVER 200)	21.4	2.8	**	0.35	0.05
GENDER					
GIRL	-10.1	1.9	**	-0.17	-0.03
SAFETY					
MEDIUM	-0.7	2.1	-	-0.01	-0.03
LOW	-14.8	2.6	**	-0.24	-0.04
OWN BEDROOM					
NO	-8.2	2.3	**	-0.14	-0.04
MOBILE PHONE					
NO	18.1	2.7	**	0.30	0.05
OUT-OF-SCHOOL ACTIVITIES					
(SCLSOC-gm)^1	0.8	2.6	-	0.03	0.09
(SCLSOC-gm)^2	-7.3	1.9	**	-0.24	-0.06
(SCLSOC-gm)^3	-4.2	1.8	*	-0.14	-0.06
(SCLIND-gm)^1	-2.2	1.8	-	-0.07	-0.06
(SCLIND-gm)^2	-9.1	1.8	**	-0.30	-0.06
(SCLPD-gm)^1	-0.4	2.2	-	-0.01	-0.07
(SCLPD-gm)^2	-10.4	2.1	**	-0.34	-0.07

Note: ** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

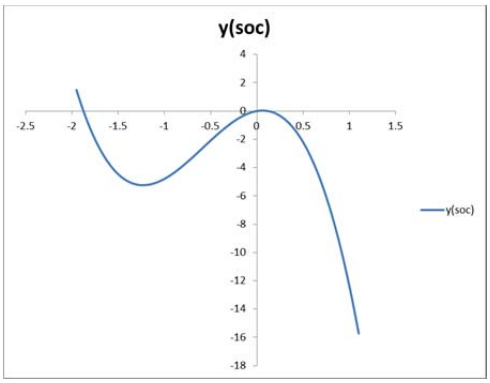
Table 5.1-2: Modelling all predictor and control variables (G4 Science - Predictors)

ASSSCIPV	ASSCI_ALL_COMBINED_MODEL_IGLS				
Response	Coef.	S.E.	Sig	effect size	+/- error
EXPLAIN SCIENCE STUDIED					
AS4SWESS_R:A FEW TIMES A YEAR	-1.3	3.2	-	-0.02	-0.05
AS4SWESS_R:ONCE OR TWICE A MONTH	0.0	3.3	-	0.00	-0.05
AS4SWESS_R:AT LEAST ONCE A WEEK	-2.6	3.6	-	-0.04	-0.06
MEMORISE SCIENCE FACTS					
AS4SMESF_R:A FEW TIMES A YEAR	4.5	3.2	-	0.07	0.05
AS4SMESF_R:ONCE OR TWICE A MONTH	13.6	3.2	**	0.22	0.05
AS4SMESF_R:AT LEAST ONCE A WEEK	24.7	3.2	**	0.41	0.05
WATCH TEACHER DEMONSTRATE EXPERIMENT					
AS4SWATE_R:A FEW TIMES A YEAR	14.3	3.3	**	0.24	0.05
AS4SWATE_R:ONCE OR TWICE A MONTH	3.3	3.4	-	0.06	0.06
AS4SWATE_R:AT LEAST ONCE A WEEK	-8.0	3.6	*	-0.13	-0.06
PLAN EXPERIMENT OR INVESTIGATION					
AS4SHPEX_R:A FEW TIMES A YEAR	6.3	2.7	*	0.10	0.04
AS4SHPEX_R:ONCE OR TWICE A MONTH	0.6	3.0	-	0.01	0.05
AS4SHPEX_R:AT LEAST ONCE A WEEK	-0.7	3.7	-	-0.01	-0.06
DO EXPERIMENT OR INVESTIGATION					
AS4SDESI_R:A FEW TIMES A YEAR	12.5	3.1	**	0.21	0.05
AS4SDESI_R:ONCE OR TWICE A MONTH	15.6	3.4	**	0.26	0.06
AS4SDESI_R:AT LEAST ONCE A WEEK	10.7	3.8	**	0.18	0.06
WORK SCIENCE PROBLEMS ON OWN					
AS4SWSPO_R:A FEW TIMES A YEAR	9.2	4.0	*	0.15	0.07
AS4SWSPO_R:ONCE OR TWICE A MONTH	12.6	3.7	**	0.21	0.06
AS4SWSPO_R:AT LEAST ONCE A WEEK	15.1	3.6	**	0.25	0.06
WORK IN GROUP EXPERIMENT OF INVESTIGATION					
AS4SHWGX:A FEW TIMES A YEAR	2.8	3.8	-	0.05	0.06
AS4SHWGX:ONCE OR TWICE A MONTH	0.2	3.8	-	0.00	0.06
AS4SHWGX:AT LEAST ONCE A WEEK	-5.3	4.0	-	-0.09	-0.07
COMPUTER IN SCIENCE LESSONS					
AS4SCOSL_R:SOME LESSONS	0.5	2.6	-	0.01	0.04
AS4SCOSL_R:ABOUT HALF THE LESSONS	-8.5	2.9	**	-0.14	-0.05
AS4SCOSL_R:EVERY OR ALMOST EVERY LESSON	-20.2	3.6	**	-0.33	-0.06

Note: ** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

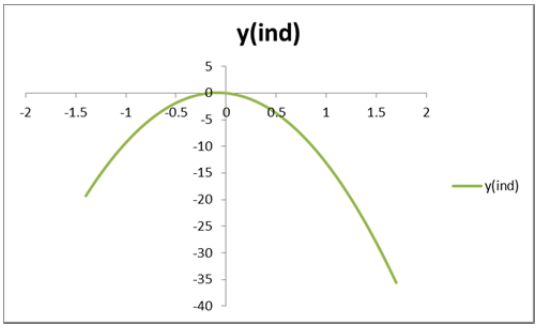
The out-of-school interests are modelled as polynomial functions and as such will require careful interpretation. Figure 5.1-1 illustrates the effect low or high responses on the Social dimension (play or talk with friends; play sports) where high responses have a negative association with science achievement.

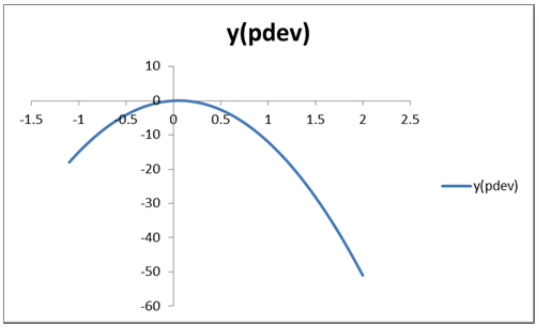
Figure 5.1-1: Out-of-school interests (Social dimension modelled as a cubic function)



The normalised scores for the other two dimensions are presented graphically in Figure 5.1-2 against the predicted change in science achievement score. The domain for each graph has been set to reflect values that contribute to the normalised function. The magnitude of reported effect sizes is an indicator of overall effect within the quadratic model. Suffice to note that the three dimensions provide significant explanation of variance in the model and that the effect for low or high responses can be interpreted from the graphical representations of each contributing dimension.

Figure 5.1-2: Out-of-school interests against G4 science achievement scores





Appendix 5.2. Model of mathematics achievement (G4) with ALL control and predictor variables

Table 5.2-1: Modelling all predictor and control variables (G4 Mathematics - Controls)

ASMMATPV	AMATH_ALL_COMBINED_MODEL_IGLS				
Response	Coef.	S.E.	Sig	effect size	+/- error
FORMATIVE YEARS IN COUNTRY(UK)					
ARRIVED IN UK AGED 5 OR OVER	-35.8	4.9	**	-0.56	-0.08
OWN LANGUAGE					
ALMOST ALWAYS	15.0	2.8	**	0.24	0.05
SOMETIMES	-11.7	4.1	**	-0.18	-0.07
NEVER	-18.9	7.4	*	-0.30	-0.12
STUDY TOOLS					
ONE OF CALC, DICT, ENCYCLOPAEDIA	2.9	7.9	-	0.05	0.12
TWO OF CALC, DICT, ENCYCLOPAEDIA	7.9	7.4	-	0.12	0.12
ALL THREE OF CALC, DICT, ENCYCLOPAEDIA	12.8	7.3	-	0.20	0.12
HOME ICT					
COMPUTER WITHOUT INTERNET	2.9	5.3	-	0.05	0.08
COMPUTER WITH INTERNET	21.9	4.5	**	0.34	0.07
BOOKS AT HOME					
NONE OR VERY FEW (0 TO 10 BOOKS)	-33.1	3.7	**	-0.52	-0.06
ONE SHELF (11 TO 25 BOOKS)	-18.8	2.8	**	-0.30	-0.04
TWO BOOKCASES (101 TO 200 BOOKS)	10.6	2.7	**	0.17	0.04
THREE OR MORE BOOKCASES (OVER 200)	14.8	2.9	**	0.23	0.05
GENDER					
GIRL	-20.4	2.0	**	-0.32	-0.03
SAFETY					
MEDIUM	0.0	2.1	-	0.00	-0.03
LOW	-12.0	2.7	**	-0.19	-0.04
OWN BEDROOM					
NO	-2.5	2.3	-	-0.04	-0.04
MOBILE PHONE					
NO	11.1	2.8	**	0.18	0.04
OUT-OF-SCHOOL ACTIVITIES					
(SCLSOC-gm)^1	4.1	2.6	-	0.13	0.08
(SCLSOC-gm)^2	-8.7	1.9	**	-0.28	-0.06
(SCLSOC-gm)^3	-4.4	1.8	*	-0.14	-0.06
(SCLIND-gm)^1	1.4	1.8	-	0.04	0.06
(SCLIND-gm)^2	-11.3	1.8	**	-0.36	-0.06
(SCLPD-gm)^1	-1.8	2.2	-	-0.06	-0.07
(SCLPD-gm)^2	-7.4	2.2	**	-0.23	-0.07

Note: ** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.2-2: Modelling all predictor and control variables (G4 Mathematics - Predictors)

ASMMATPV	AMATH_ALL_COMBINED_MODEL_IGLS				
Response	Coef.	S.E.	Sig	effect size	+/- error
EXPLAIN ANSWERS (NEVER)					
AS4MHEXP_R:SOME LESSONS	9.5	3.3	**	0.15	0.05
AS4MHEXP_R:ABOUT HALF THE LESSONS	14.0	3.6	**	0.22	0.06
AS4MHEXP_R:EVERY OR ALMOST EVERY LESSON	11.0	3.7	**	0.17	0.06
MEMORISE PROCEDURES ETC. (NEVER)					
AS4MHMWP_R:SOME LESSONS	3.8	4.3	-	0.06	0.07
AS4MHMWP_R:ABOUT HALF THE LESSONS	7.3	4.4	-	0.11	0.07
AS4MHMWP_R:EVERY OR ALMOST EVERY LESSON	19.1	4.4	**	0.30	0.07
MEASURE THINGS IN CLASS (NEVER)					
AS4MHMCL_R:SOME LESSONS	1.9	3.0	-	0.03	0.05
AS4MHMCL_R:ABOUT HALF THE LESSONS	-16.8	4.2	**	-0.26	-0.07
AS4MHMCL_R:EVERY OR ALMOST EVERY LESSON	-28.1	4.7	**	-0.44	-0.08
MAKE TABLES CHARTS AND GRAPHS (NEVER)					
AS4MHTCG_R:SOME LESSONS	24.9	4.0	**	0.39	0.06
AS4MHTCG_R:ABOUT HALF THE LESSONS	20.1	4.3	**	0.32	0.07
AS4MHTCG_R:EVERY OR ALMOST EVERY LESSON	13.8	4.7	**	0.22	0.07
WORK ON OWN (NEVER)					
AS4MHWPO_R:SOME LESSONS	24.6	7.5	**	0.39	0.12
AS4MHWPO_R:ABOUT HALF THE LESSONS	29.8	7.5	**	0.47	0.12
AS4MHWPO_R:EVERY OR ALMOST EVERY LESSON	41.9	7.4	**	0.66	0.12
WORK IN GROUPS (NEVER)					
AS4MHWSG_R:SOME LESSONS	11.0	4.1	**	0.17	0.06
AS4MHWSG_R:ABOUT HALF THE LESSONS	14.8	4.5	**	0.23	0.07
AS4MHWSG_R:EVERY OR ALMOST EVERY LESSON	-1.2	4.5	-	-0.02	-0.07
USE A CALCULATOR IN CLASS (NEVER)					
AS4MHCAL_R:SOME LESSONS	12.4	2.3	**	0.20	0.04
AS4MHCAL_R:ABOUT HALF THE LESSONS	4.4	4.5	-	0.07	0.07
AS4MHCAL_R:EVERY OR ALMOST EVERY LESSON	-21.7	7.4	**	-0.34	-0.12
USE A COMPUTER IN CLASS					
AS4MHCOM_R:SOME LESSONS	-3.1	2.3	-	-0.05	-0.04
AS4MHCOM_R:ABOUT HALF THE LESSONS	-13.0	3.6	**	-0.20	-0.06
AS4MHCOM_R:EVERY OR ALMOST EVERY LESSON	-29.9	4.4	**	-0.47	-0.07

Note: ** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Appendix 5.3. G8 Summaries: Science

Appendix 5.3.1. Models of G8 Science Achievement

Table 5.3-1: Explain My Studies (Model C) compared with Unconditional (Model A)

MCMC Estimation (220,000 iterations)		*Unconditional Model (Model A)*		*Explain my studies (Model C)*		
Response: BSSSCIPV		Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part						
CONSTANT	β_0	495.7	4.0	477.6	9.2	**
EXPLAIN MY STUDIES (NEVER)						
SOME LESSONS	β_{33}			2.4	3.9	-
ABOUT HALF THE LESSONS	β_{34}			11.4	3.9	**
EVERY OR ALMOST EVERY LESSON	β_{35}			13.6	3.9	**
Random Part						*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	2150.1	313.5	885.2	143.0	0.59
Level-3: IDTEACH	v_{0kl}	491.0	73.1	279.0	53.9	0.43
Level-2: IDSTUD	u_{0jkl}	3254.5	87.2	2414.2	66.6	0.26
Level-1: IDCASE	e_{0ijkl}	704.4	8.1	704.4	8.2	0.00
Total Variance		6600.0		4282.7		0.35
DIC:		177609.9		177563.3		
pD:		3566.6		3520.1		

Note: N=18525 cases, 3668 students, with 773 teachers, in 129 schools;
** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.3-2: Watch teacher demonstrate an experiment (Model D) compared with Model A

MCMC Estimation (450,000 iterations)		*Unconditional Model (Model A)*		*Watch teacher demonstrate (Model D)*		
Response: BSSSCIPV		Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part						
CONSTANT	β_0	495.0	4.2	471.6	10.6	**
WATCH TEACHER DEMONSTRATE (NEVER)						
SOME LESSONS	β_{33}			11.0	6.4	-
ABOUT HALF THE LESSONS	β_{34}			20.7	6.4	**
EVERY OR ALMOST EVERY LESSON	β_{35}			13.4	6.3	*
Random Part						*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	2162.3	315.8	881.0	143.2	0.59
Level-3: IDTEACH	v_{0kl}	487.1	74.0	281.8	52.9	0.42
Level-2: IDSTUD	u_{0jkl}	3287.7	86.39	2444.1	66.7	0.26
Level-1: IDCASE	e_{0ijkl}	705.6	8.1	705.7	8.1	0.00
Total Variance		6642.7		4312.6		0.35
DIC:		179657.1		179612.0 (45.1)		
pD:		3606.7		3561.1 (45.7)		

Note: N=18735 cases, 3747 students, with 774 teachers, in 129 schools;
** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.3-3: Design or plan a science experiment (Model E) compared with Model A

MCMC Estimation (150,000 iterations)		*Unconditional Model (Model A)*		*Design or plan an experiment (Model E)*		
Response: BSSSCIPV		Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part						
CONSTANT	β_0	494.9	4.5	480.5	9.2	**
DESIGN OR PLAN AN EXPERIMENT (NEVER)						
SOME LESSONS	β_{33}			6.9	4.0	-
ABOUT HALF THE LESSONS	β_{34}			10.9	3.9	**
EVERY OR ALMOST EVERY LESSON	β_{35}			5.1	4.0	-
Random Part						*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	2147.7	316.0	868.9	142.7	0.60
Level-3: IDTEACH	v_{0kl}	496.4	76.4	288.2	53.6	0.42
Level-2: IDSTUD	u_{0jkl}	3292.4	88.5	2458.3	67.2	0.25
Level-1: IDCASE	e_{0ijkl}	703.8	8.2	703.7	8.2	0.00
Total Variance		6640.3		4319.2		0.35
DIC:		178268.0		178223.2	(44.6)	
pD:		3581.2		3536.6	(44.6)	

Note: N=18595 cases, 3719 students, with 773 teachers, in 129 schools;
** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.3-4: Conduct an experiment or investigation (Model F) compared with Model A

MCMC Estimation (150,000 iterations)		*Unconditional Model (Model A)*		*Conduct an experiment (Model F)*		
Response: BSSSCIPV		Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part						
CONSTANT	β_0	494.3	4.1	470.3	10.4	**
CONDUCT AN EXPERIMENT (NEVER)						
SOME LESSONS	β_{33}			13.2	6.5	*
ABOUT HALF THE LESSONS	β_{34}			22.9	6.5	**
EVERY OR ALMOST EVERY LESSON	β_{35}			21.2	6.4	**
Random Part						*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	2148.9	312.5	883.6	144.0	0.59
Level-3: IDTEACH	v_{0kl}	493.1	75.7	290.0	53.8	0.41
Level-2: IDSTUD	u_{0jkl}	3265.4	87.3	2429.3	67.1	0.26
Level-1: IDCASE	e_{0ijkl}	705.1	8.2	705.1	8.2	0.00
Total Variance		6612.6		4308.1		0.35
DIC:		177151.0		177105.9	(45.2)	
pD:		3556.7		3511.3	(45.4)	

Note: N=18475 cases, 3695students, with 773 teachers, in 129 schools;
** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.3-5: Work on own (Model G) compared with Model A

MCMC Estimation (200,000 iterations)		*Unconditional Model (Model A)*		*Work on own (Model G)*			
Response: BSSSCIPV		Coefficient	s.e.	Coef.	s.e.	Sig.	
Fixed Part							
CONSTANT	β_0	494.4	4.3	479.1	9.0	**	
WORK ON OWN (NEVER)							
SOME LESSONS	β_{33}			4.5	3.5	-	
ABOUT HALF THE LESSONS	β_{34}			12.0	3.5	**	
EVERY OR ALMOST EVERY LESSON	β_{35}			11.1	3.8	**	
Random Part							*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	2161.2	315.0	897.9	145.0		0.58
Level-3: IDTEACH	v_{0kl}	490.4	75.2	286.8	53.7		0.42
Level-2: IDSTUD	u_{0jkl}	3247.7	87.6	2437.1	67.0		0.26
Level-1: IDCASE	e_{0ijkl}	704.2	8.2	704.3	8.2		0.00
Total Variance		6630.6		4326.0			0.35
DIC:		178805.8		178761.3	(44.5)		
pD:		3590.9		3545.6	(45.3)		

Note: N=18650 cases, 3730 students, with 774 teachers, in 129 schools;
** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.3-6: Work in small groups (Model H) compared with Model A

MCMC Estimation (200,000 iterations)		*Unconditional Model (Model A)*		*Work in small groups (Model H)*			
Response: BSSSCIPV		Coefficient	s.e.	Coef.	s.e.	Sig.	
Fixed Part							
CONSTANT	β_0	494.6	4.3	472.9	10.9	**	
WORK IN SMALL GROUPS (NEVER)							
SOME LESSONS	β_{33}			9.3	7.4	-	
ABOUT HALF THE LESSONS	β_{34}			21.0	7.4	**	
EVERY OR ALMOST EVERY LESSON	β_{35}			19.5	7.3	**	
Random Part							*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	2181.9	317.6	887.8	143.8		0.59
Level-3: IDTEACH	v_{0kl}	485.8	74.8	273.0	52.4		0.44
Level-2: IDSTUD	u_{0jkl}	3289.3	87.7	2450.2	67.2		0.26
Level-1: IDCASE	e_{0ijkl}	703.9	8.1	703.9	8.1		0.00
Total Variance		6660.9		4314.8			0.35
DIC:		178797.1		178751.2	(45.9)		
pD:		3591.7		3545.9	(45.8)		

Note: N=18650 cases, 3730 students, with 773 teachers, in 129 schools;
** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.3-7: Relate learning to Daily Life (Model I) compared with Model A

MCMC Estimation (250,000 iterations)		*Unconditional Model (Model A)*		*Daily Life (Model I)*			
Response: BSSSCIPV		Coefficient	s.e.	Coef.	s.e.	Sig.	
Fixed Part							
CONSTANT	β_0	495.3	4.3	485.6	8.8	**	
DAILY LIFE (NEVER)							
SOME LESSONS	β_{33}			-0.7	2.8	-	
ABOUT HALF THE LESSONS	β_{34}			4.7	2.8	-	
EVERY OR ALMOST EVERY LESSON	β_{35}			0.6	3.2	-	
Random Part							*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	2183.4	320.8	922.4	149.0		0.58
Level-3: IDTEACH	v_{0kl}	483.8	74.8	276.5	52.8		0.43
Level-2: IDSTUD	u_{0jkl}	3277.5	88.0	2449.4	67.7		0.25
Level-1: IDCASE	e_{0ijkl}	705.8	8.2	705.8	8.3		0.00
Total Variance		6650.4		4354.0			0.35
DIC:		177839.0		177797.3	(41.7)		
pD:		3570.0		3526.6	(43.4)		

Note: N=18545 cases, 3709 students, with 774 teachers, in 129 schools;
** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.3-8: Review homework (Model J) compared with Model A

MCMC Estimation (174,000 iterations)		*Unconditional Model (Model A)*		*Review Homework (Model J)*			
Response: BSSSCIPV		Coefficient	s.e.	Coef.	s.e.	Sig.	
Fixed Part							
CONSTANT	β_0	494.8	4.3	491.4	8.9	**	
REVIEW HOMEWORK (NEVER)							
SOME LESSONS	β_{33}			-2.8	2.4	-	
ABOUT HALF THE LESSONS	β_{34}			-6.4	2.7	*	
EVERY OR ALMOST EVERY LESSON	β_{35}			-12.0	2.8	**	
Random Part							*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	2171.1	317.0	869.7	141.5		0.60
Level-3: IDTEACH	v_{0kl}	477.8	74.4	283.0	53.6		0.41
Level-2: IDSTUD	u_{0jkl}	3286.0	88.2	2440.7	67.3		0.26
Level-1: IDCASE	e_{0ijkl}	706.5	8.2	706.5	8.3		0.00
Total Variance		6641.2		4299.9			0.35
DIC:		177522.7		177475.9	(46.8)		
pD:		3564.0		3517.8	(46.2)		

Note: N=18510 cases, 3702 students, with 773 teachers, in 129 schools;
** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.3-9: Quiz or test (Model K) compared with Model A

MCMC Estimation (321,000 iterations)		*Unconditional Model (Model A)*		*Quiz or Test (Model K)*		
Response: BSSSCIPV		Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part						
CONSTANT	β_0	494.5	4.2	476.1	9.9	**
QUIZ OR TEST (NEVER)						
SOME LESSONS	β_{33}			14.1	5.6	*
ABOUT HALF THE LESSONS	β_{34}			4.4	5.9	-
EVERY OR ALMOST EVERY LESSON	β_{35}			-12.1	6.2	-
Random Part						*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	2158.7	313.8	825.9	135.8	0.62
Level-3: IDTEACH	v_{0kl}	486.6	75.9	280.5	53.2	0.42
Level-2: IDSTUD	u_{0jkl}	3285.2	87.6	2416.5	66.7	0.26
Level-1: IDCASE	e_{0ijkl}	706.1	8.2	706.2	8.2	0.00
Total Variance		6636.6		4229.2		0.36
DIC:		179048.3		179000.6	(47.7)	
pD:		3594.7		3546.8	(47.9)	

Note: N=18670 cases, 3734 students, with 774 teachers, in 129 schools;
** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.3-10: Lecture-style presentation (Model L) compared with Model A

MCMC Estimation (340,000 iterations)		*Unconditional Model (Model A)*		*Lecture-style presentation (Model L)*		
Response: BSSSCIPV		Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part						
CONSTANT	β_0	494.8	4.2	474.4	10.1	**
LECTURE-STYLE PRESENTATION (NEVER)						
SOME LESSONS	β_{33}			2.0	6.0	-
ABOUT HALF THE LESSONS	β_{34}			11.7	5.8	*
EVERY OR ALMOST EVERY LESSON	β_{35}			18.5	5.7	**
Random Part						*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	2160.7	316.8	860.6	139.4	0.60
Level-3: IDTEACH	v_{0kl}	487.2	74.8	254.8	51.2	0.48
Level-2: IDSTUD	u_{0jkl}	3272.0	86.9	2420.0	66.7	0.26
Level-1: IDCASE	e_{0ijkl}	705.1	8.2	705.0	8.2	0.00
Total Variance		6625.0		4240.6		0.36
DIC:		178206.4		178159.9	(46.5)	
pD:		3578.2		3531.2	(47.0)	

Note: N=18585 cases, 3717 students, with 774 teachers, in 129 schools;
** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.3-11: Computer in science lesson (Model M) compared with Model A

MCMC Estimation (200,000 iterations)		*Unconditional Model (Model A)*		*Computer (Model M)*		
Response: BSSSCIPV		Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part						
CONSTANT	β_0	494.9	4.2	489.4	8.7	**
COMPUTER (NEVER)						
SOME LESSONS	β_{33}			2.5	2.1	-
ABOUT HALF THE LESSONS	β_{34}			-12.9	3.6	**
EVERY OR ALMOST EVERY LESSON	β_{35}			-31.8	5.1	**
Random Part						*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	2174.9	316.3	817.3	135.2	0.62
Level-3: IDTEACH	v_{0kl}	479.1	74.7	277.2	52.3	0.42
Level-2: IDSTUD	u_{0jkl}	3288.4	88.0	2427.6	66.3	0.26
Level-1: IDCASE	e_{0ijkl}	706.0	8.2	706.0	8.1	0.00
Total Variance		6648.5		4228.1		0.36
DIC:		179332.3		179286.7	(45.6)	
pD:		3600.4		3554.0	(46.4)	

Note: N=18700 cases, 3740 students, with 774 teachers, in 129 schools;
** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.3-12: Computer for school work (Model N) compared with Model A

MCMC Estimation (149,000 iterations)		*Unconditional Model (Model A)*		*Computer for School Work (Model N)*		
Response: BSSSCIPV		Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part						
CONSTANT	β_0	495.5	4.4	480.0	8.9	**
COMPUTER FOR SCHOOL WORK (NO TIME)						
A FEW TIMES A YEAR	β_{33}			12.1	2.3	**
ONCE OR TWICE A MONTH	β_{34}			6.3	2.4	**
AT LEAST ONCE A WEEK	β_{35}			-8.0	3.4	*
EVERY DAY	β_{36}			-19.8	6.9	**
Random Part						*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	2150.3	312.4	830.2	136.4	0.61
Level-3: IDTEACH	v_{0kl}	479.1	76.3	262.0	53.2	0.45
Level-2: IDSTUD	u_{0jkl}	3263.6	88.0	2432.1	68.0	0.26
Level-1: IDCASE	e_{0ijkl}	706.7	8.3	706.8	8.2	0.00
Total Variance		6605.3		4231.1		0.36
DIC:		175660.7		175615.2	(45.5)	
pD:		3525.7		3480.0	(45.6)	

Note: N=18315 cases, 3663 students, with 769 teachers, in 129 schools;
** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.3-13: Combination of 'CONDUCT', 'WATCH' and 'PLAN' (including control variables)

IGLS Estimation	CONDUCT + WATCH (Model O)		CONDUCT + WATCH + PLAN (Model P)	
Response: BSSSCIPV	Coef.(s.e.)	Sig.	Coef.(s.e.)	Sig.
Fixed Part				
CONSTANT β_0	464.7 (11.3)	**	464.2 (11.3)	**
CONDUCT EXPERIMENTS (NEVER)				
SOME LESSONS	11.9 (6.9)	-	11.1 (7.0)	-
ABOUT HALF THE LESSONS	19.7 (6.8)	**	18.9 (7.0)	**
EVERY OR ALMOST EVERY LESSON	19.8 (6.9)	**	21.0 (7.1)	**
WATCH TEACHER DEMONSTRATE EXPERIMENT (NEVER)				
SOME LESSONS	5.6 (6.9)	-	4.3 (7.0)	-
ABOUT HALF THE LESSONS	12.6 (6.9)	-	11.4 (7.0)	-
EVERY OR ALMOST EVERY LESSON	4.6 (6.9)	-	4.6 (7.1)	-
PLAN SCIENCE EXPERIMENT OR INVESTIGATION (NEVER)				
SOME LESSONS			3.3 (4.1)	-
ABOUT HALF THE LESSONS			4.0 (4.1)	-
EVERY OR ALMOST EVERY LESSON			-2.5 (4.4)	-
Random Part		*Total Variance Reduction (%)*		*Total Variance Reduction (%)*
Level-4: IDSCHOOL f_{0l}	836.0 (126.5)	0.60	828.7 (125.6)	0.60
Level-3: IDTEACH v_{0kl}	293.8 (51.2)	0.41	295.0 (51.2)	0.41
Level-2: IDSTUD u_{0jkl}	2389.8 (65.3)	0.27	2384.3 (65.2)	0.27
Level-1: IDCASE e_{0ijkl}	703.8 (8.2)	0.00	703.8 (8.2)	0.0
Total Variance	4223.4	0.35	4211.8	0.35
*-2*loglikelihood (difference from previous model)*	182985.6	(16.9)	182978.2	(7.4)

Note: N=17430 cases, 3564 students, with 250 teachers, in 137 schools;
** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.3-14: Science combination of 'explain' and 'memorise' (including control variables)

IGLS Estimation	EXPLAIN MY STUDIES (Model B′)		MEMORISE SCIENCE FACTS & PRINCIPLES (Model C′)		EXPLAIN + MEMORISE (Model Q)	
Response: BSSSCIPV	Coef.(s.e.)	Sig.	Coef.(s.e.)	Sig.	Coef. (s.e.)	Sig.
Fixed Part						
CONSTANT	476.1 (9.3)	**	474.2 (9.2)	**	471.3 (9.5)	**
EXPLAIN (NEVER)						
SOME LESSONS	2.8 (3.9)	-			0.8 (4.0)	-
ABOUT HALF THE LESSONS	11.5 (3.9)	**			8.4 (4.0)	*
EVERY OR ALMOST EVERY LESSON	13.7 (3.9)	**			11.1 (4.2)	-
MEMORISE (NEVER)						
SOME LESSONS			8.6 (3.6)	**	6.9 (3.7)	-
ABOUT HALF THE LESSONS			15.4 (3.6)	**	11.2 (3.8)	**
EVERY OR ALMOST EVERY LESSON			11.9 (3.8)	**	6.0 (4.0)	-
Random Part		*Total Variance Reduction*		*Total Variance Reduction*		*Total Variance Reduction*
Level-4: IDSCHOOL f_{0l}	850.5 (128.0)	0.60	838.5 (126.5)	0.61	834.8 (125.9)	0.61
Level-3: IDTEACH v_{0kl}	285.8 (50.3)	0.42	286.3 (50.4)	0.42	284.2 (50.1)	0.42
Level-2: IDSTUD u_{0jkl}	2378.2 (64.9)	0.27	2385.8 (65.1)	0.27	2372.0 (64.7)	0.27
Level-1: IDCASE e_{0ijkl}	704.1 (8.2)	0.00	704.1 (8.2)	0.00	704.1 (8.2)	0.00
Total Variance	4218.5	0.36	4214.7.0	0.26	4195.0	0.36
-2*loglikelihood:	183672.0		183681.1		183660.6	(11.4)

Note: N=18365 cases, 3673 students, with 771 teachers, in 129 schools;
** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Appendix 5.3.2. Model of science achievement (G8) with ALL control and predictor variables

Table 5.3-15: Modelling all predictor and control variables (G8 Science - Controls)

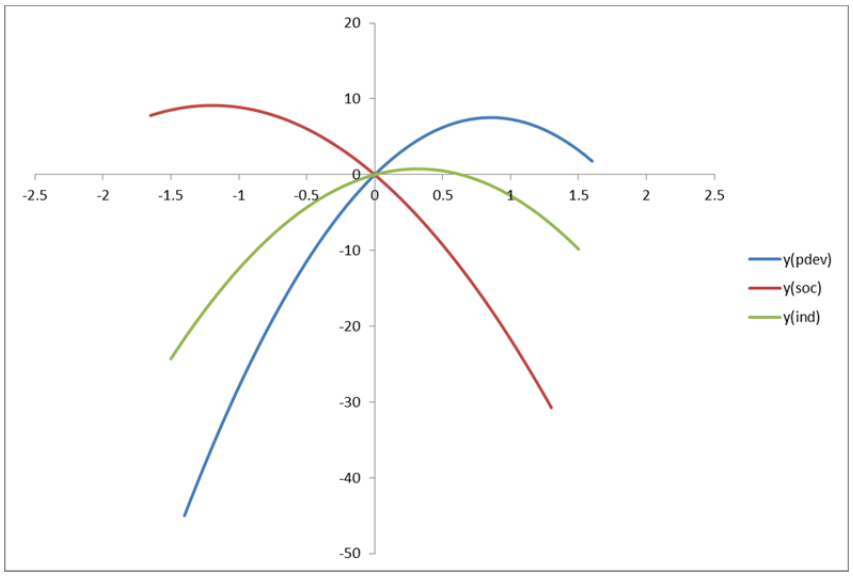
BSSSCIPV	BSCI_ALL_COMBINED_MODEL_IGLS				
Response	Coef.	S.E.	Sig	effect size	+/- error
OWN LANGUAGE					
ALMOST ALWAYS	-4.8	4.0	-	-0.08	0.07
SOMETIMES	-5.8	5.7	-	-0.10	0.10
NEVER	-34.7	10.0	**	-0.61	0.18
HOME ICT					
COMPUTER WITHOUT INTERNET	7.0	7.1	-	0.12	0.12
HOME COMPUTER AND INTERNET	21.9	6.2	**	0.39	0.11
BOOKS AT HOME					
NONE OR VERY FEW (0 TO 10 BOOKS)	-30.8	3.0	**	-0.54	0.05
ONE SHELF (11 TO 25 BOOKS)	-12.0	2.6	**	-0.21	0.05
TWO BOOKCASES (101 TO 200 BOOKS)	13.5	2.9	**	0.24	0.05
THREE OR MORE BOOKCASES (OVER 200)	35.9	2.9	**	0.64	0.06
OWN BEDROOM/ STUDY SPACE					
BEDROOM_NO	-5.7	2.6	*	-0.10	0.05
WORK IN PAID JOB					
WKPJ_LESS THAN 1 HOUR	-12.7	3.4	**	-0.23	0.06
WKPJ_1 TO 2 HOURS	-7.7	3.6	*	-0.14	0.06
WKPJ_MORE THAN 2 BUT LESS THAN 4 HRS	-6.3	4.4	-	-0.11	0.08
WKPJ_4 OR MORE HOURS	-5.5	4.4	-	-0.10	0.08
PLAY SPORTS					
PLSP_LESS THAN 1 HOUR	1.8	3.1	-	0.03	0.06
PLSP_1 TO 2 HOURS	5.7	3.0	-	0.10	0.05
PLSP_MORE THAN 2 BUT LESS THAN 4 HOURS	2.1	3.3	-	0.04	0.06
PLSP_4 OR MORE HOURS	-9.4	3.4	**	-0.17	0.06
GENDER					
GIRL	-15.5	2.1	**	-0.55	0.07
FORMATIVE YEARS IN COUNTRY(UK)					
ARRIVED IN UK AGED 5 TO 10 YEARS	-10.1	8.6	-	-0.36	0.30
ARRIVED IN UK OLDER THAN 10 YEARS	-30.8	6.8	**	-1.09	0.24
STUDY TOOLS					
ONE OF CALC, DICT, ENCYCLOPAEDIA	6.0	6.1	-	0.21	0.22
TWO OF CALC, DICT, ENCYCLOPAEDIA	12.6	5.5	*	0.45	0.20
ALL THREE OF CALC, DICT, ENCYCLOPAEDIA	16.9	5.5	**	0.60	0.20

BSSSCIPV	BSCI_ALL_COMBINED_MODEL_IGLS				
Response	Coef.	S.E.	Sig	effect size	+/- error
OUT-OF-SCHOOL ACTIVITIES					
(BSCLSOC-gm)^1	-15.3	1.8	**	-0.54	0.07
(BSCLSOC-gm)^2	-6.4	1.8	**	-0.11	0.03
(BSCLIND-gm)^1	4.8	1.5	**	0.08	0.03
(BSCLIND-gm)^2	-7.6	1.5	**	-0.13	0.03
(BSCLPD-gm)^1	17.7	2.6	**	0.31	0.05
(BSCLPD-gm)^2	-10.3	2.5	**	-0.18	0.04

Note: ** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

The out-of-school interests in G8 Science are all modelled as pseudo-continuous quadratic functions. The model coefficients permit graphs to be constructed to illustrate association with achievement scores; these are presented in Figure 5.3-1.

Figure 5.3-1: Out-of-school activities against G8 science achievement scores



Inspection of effect sizes attributed to each response in those out-of-school interests highlights the strong association with achievement scores for the Social and Personal Development dimensions ($|\Delta| = 0.66$ and $|\Delta| = 0.76$ respectively). These effects are confirmed in the graphical representation with Personal Development dimension showing the larger residuals and the Individual dimension offering the least explanation given its relatively flatter structure.

Table 5.3-16: Modelling all predictor and control variables (G8 Science - Predictors)

BSSSCIPV	BSCI_ALL_COMBINED_MODEL_IGLS				
Response	Coef.	S.E.	Sig	effect size	+/- error
BS4SHEOS_R:SOME LESSONS	3.2	4.3	-	0.06	0.08
BS4SHEOS_R:ABOUT HALF THE LESSONS	8.2	4.5	-	0.15	0.08
BS4SHEOS_R:EVERY OR ALMOST EVERY LESSON	12.7	4.6	**	0.22	0.08
BS4SHFAP_R:SOME LESSONS	3.1	4.0	-	0.05	0.07
BS4SHFAP_R:ABOUT HALF THE LESSONS	7.3	4.2	-	0.13	0.07
BS4SHFAP_R:EVERY OR ALMOST EVERY LESSON	7.0	4.4	-	0.12	0.08
BS4SHWPO_R:SOME LESSONS	2.6	3.7	-	0.05	0.06
BS4SHWPO_R:ABOUT HALF THE LESSONS	6.9	3.7	-	0.12	0.07
BS4SHWPO_R:EVERY OR ALMOST EVERY LESSON	10.9	4.0	**	0.19	0.07
BS4SHWGO_R:SOME LESSONS	4.5	9.3	-	0.08	0.16
BS4SHWGO_R:ABOUT HALF THE LESSONS	11.0	9.3	-	0.19	0.16
BS4SHWGO_R:EVERY OR ALMOST EVERY LESSON	9.5	9.3	-	0.17	0.16
BS4SHLSP_R:SOME LESSONS	1.1	6.9	-	0.02	0.12
BS4SHLSP_R:ABOUT HALF THE LESSONS	6.9	6.8	-	0.12	0.12
BS4SHLSP_R:EVERY OR ALMOST EVERY LESSON	14.3	6.7	*	0.25	0.12
BS4SHDEI_R:SOME LESSONS	-5.9	7.4	-	-0.11	0.13
BS4SHDEI_R:ABOUT HALF THE LESSONS	0.2	7.4	-	0.00	0.13
BS4SHDEI_R:EVERY OR ALMOST EVERY LESSON	-6.1	7.5	-	-0.11	0.13
BS4SHCEI_R:SOME LESSONS	5.5	8.0	-	0.10	0.14
BS4SHCEI_R:ABOUT HALF THE LESSONS	9.0	8.1	-	0.16	0.14
BS4SHCEI_R:EVERY OR ALMOST EVERY LESSON	8.9	8.2	-	0.16	0.15
BS4SHPEI_R:SOME LESSONS	1.7	4.3	-	0.03	0.08
BS4SHPEI_R:ABOUT HALF THE LESSONS	0.9	4.4	-	0.02	0.08
BS4SHPEI_R:EVERY OR ALMOST EVERY LESSON	-3.4	4.7	-	-0.06	0.08

Table 5.3-17: Modelling all predictor and control variables (G8 Science – Predictors) Continued

BSSSCIPV	BSCI_ALL_COMBINED_MODEL_IGLS				
Response	Coef.	S.E.	Sig	effect size	+/- error
BS4SHROH_R:SOME LESSONS	-5.0	2.6	-	-0.09	0.05
BS4SHROH_R:ABOUT HALF THE LESSONS	-12.7	2.9	**	-0.23	0.05
BS4SHROH_R:EVERY OR ALMOST EVERY LESSON	-15.6	3.1	**	-0.28	0.06
BS4SHHQT_R:SOME LESSONS	7.5	6.0	-	0.13	0.11
BS4SHHQT_R:ABOUT HALF THE LESSONS	-0.5	6.4	-	-0.01	0.11
BS4SHHQT_R:EVERY OR ALMOST EVERY LESSON	-14.2	6.8	*	-0.25	0.12
BS4SHCOM_R:SOME LESSONS	2.3	2.2	-	0.04	0.04
BS4SHCOM_R:ABOUT HALF THE LESSONS	-10.4	3.8	**	-0.18	0.07
BS4SHCOM_R:EVERY OR ALMOST EVERY LESSON	-20.4	5.8	**	-0.36	0.10
BS4SCSWS_R_A FEW TIMES A YEAR	8.8	2.4	**	0.16	0.04
BS4SCSWS_R_ONCE OR TWICE A MONTH	4.7	2.5	-	0.08	0.04
BS4SCSWS_R_AT LEAST ONCE A WEK	-9.7	3.6	**	-0.17	0.06
BS4SCSWS_R_EVERY DAY	-15.6	7.0	*	-0.28	0.12
BS4SHMDL_R:SOME LESSONS	-4.5	3.0	-	-0.08	0.05
BS4SHMDL_R:ABOUT HALF THE LESSONS	-2.1	3.2	-	-0.04	0.06
BS4SHMDL_R:EVERY OR ALMOST EVERY LESSON	0.1	3.6	-	0.00	0.06

Note: ** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Appendix 5.4. G8 Summaries: Mathematics

Appendix 5.4.1. Models of GR8 Mathematics Achievement

Table 5.4-1: Memorise Procedures etc. (Model B) compared with Unconditional (Model A)

MCMC Estimation		*Unconditional Model (Model A)*		*Interpret data (Model B)*		
Response: BSMMATPV		Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part						
CONSTANT	β_0	477.1	3.8	487.1	4.3	**
MEMORISE PROCEDURES (NEVER)						
SOME LESSONS	β_{30}			9.8	2.1	**
ABOUT HALF THE LESSONS	β_{31}			9.9	2.2	**
EVERY OR ALMOST EVERY LESSON	β_{32}			10.2	2.4	**
Random Part						*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	229.2	276.2	158.9	208.1	0.31
Level-3: IDTEACH	v_{0kl}	4092.7	426.4	3288.5	331.7	0.20
Level-2: IDSTUD	u_{0jkl}	1429.0	37.0	1269.5	33.3	0.11
Level-1: IDCASE	e_{0ijkl}	502.3	5.8	502.2	5.8	0.00
Total Variance		6283.6		5243.3		0.17
DIC:		172276.9		172254.4	(22.5)	
pD:		3501.9		3478.6	(23.3)	

Note: N=18635 cases, 3727 students, with 312 teachers, in 129 schools;

** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.4-2: Explain my answers (Model C) compared with Model A

MCMC Estimation		*Unconditional Model (Model A)*		*Work in small groups (Model E)*		
Response: BSMMATPV		Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part						
CONSTANT	β_0	478.3	4.2	489.0	5.1	**
EXPLAIN MY ANSWERS (NEVER)						
SOME LESSONS	β_{30}			5.8	3.3	-
ABOUT HALF THE LESSONS	β_{31}			7.9	3.3	*
EVERY OR ALMOST EVERY LESSON	β_{32}			7.6	3.1	*
Random Part						*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	219.3	276.6	44.4	139.8	0.80
Level-3: IDTEACH	v_{0kl}	4098.9	424.2	3499.5	314.0	0.15
Level-2: IDSTUD	u_{0jkl}	1421.8	37.0	1267.8	33.2	0.11
Level-1: IDCASE	e_{0ijkl}	503.3	5.8	503.3	5.8	0.00
Total Variance		6243.3		5315.0		0.15
DIC:		172219.2		172192.6	(26.6)	
pD:		3498.9		3473.4	(25.5)	

Note: N=18615 cases, 3723 students, with 311 teachers, in 129 schools;

** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.4-3: Interpret Data (Model D) compared with Model A

MCMC Estimation		*Unconditional Model (Model A)*		*Interpret data (Model D)*		
Response: BSMMATPV		Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part						
CONSTANT	β_0	477.8	3.9	493.5	5.4	**
INTERPRET DATA (NEVER)						
SOME LESSONS	β_{30}			4.6	3.3	-
ABOUT HALF THE LESSONS	β_{31}			-2.9	3.5	-
EVERY OR ALMOST EVERY LESSON	β_{32}			-10.6	3.9	**
Random Part						*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	98.0	208.7	46.3	111.0	0.53
Level-3: IDTEACH	v_{0kl}	4257.4	411.9	3436.8	313.3	0.19
Level-2: IDSTUD	u_{0jkl}	1425.6	36.8	1257.7	33.0	0.12
Level-1: IDCASE	e_{0ijkl}	502.6	5.8	502.6	5.8	0.00
Total Variance		6283.6		5243.3		0.17
DIC:		176153.3		175611.4	(541.9)	
pD:		3444.4		3427.7	(9.8)	

Note: N=18670 cases, 3734 students, with 312 teachers, in 129 schools;

** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.4-4: Work in small groups (Model E) compared with Model A

MCMC Estimation		*Unconditional Model (Model A)*		*Work in small groups (Model E)*		
Response: BSMMATPV		Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part						
CONSTANT	β_0	477.5	4.2	496.7	4.4	**
WORK IN SMALL GROUPS (NEVER)						
SOME LESSONS	β_{30}			-1.2	1.6	-
ABOUT HALF THE LESSONS	β_{31}			-3.2	2.2	-
EVERY OR ALMOST EVERY LESSON	β_{32}			-15.0	2.7	**
Random Part						*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	280.6	362.4	3.6	13.7	0.99
Level-3: IDTEACH	v_{0kl}	4111.5	451.4	3519.6	303.0	0.14
Level-2: IDSTUD	u_{0jkl}	1414.4	36.7	1256.6	33.4	0.11
Level-1: IDCASE	e_{0ijkl}	503.2	5.8	503.2	5.8	0.00
Total Variance		6309.7		5283.1		0.16
DIC:		172219.2		172192.6	(26.6)	
pD:		3498.9		3473.4	(25.5)	

Note: N=18625 cases, 3725 students, with 312 teachers, in 129 schools;

** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.4-5: Work on own (Model F) compared with Model A

MCMC Estimation		*Unconditional Model (Model A)*		*Work on own (Model F)*		
Response: BSMMATPV		Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part						
CONSTANT	β_0	477.5	3.8	487.5	5.3	**
WORK ON OWN (NEVER)						
SOME LESSONS	β_{30}			3.0	3.3	-
ABOUT HALF THE LESSONS	β_{31}			5.7	3.3	-
EVERY OR ALMOST EVERY LESSON	β_{32}			12.5	3.3	**
Random Part						*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	160.4	244.1	53.3	98.4	0.67
Level-3: IDTEACH	v_{0kl}	4183.0	417.6	3437.6	307.0	0.18
Level-2: IDSTUD	u_{0jkl}	1426.7	37.0	1263.5	33.4	0.11
Level-1: IDCASE	e_{0ijkl}	502.3	5.8	502.4	5.8	0.00
Total Variance		6272.4		5256.8		0.16
DIC:		171862.5		171838.6	(23.9)	
pD:		3492.9		3468.3	(24.7)	

Note: N=18590 cases, 3718 students, with 312 teachers, in 129 schools;
** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.4-6: Daily Life (Model G) compared with Model A

MCMC Estimation		*Unconditional Model (Model A)*		*Daily Life (Model G)*		
Response: BSMMATPV		Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part						
CONSTANT	β_0	477.8	3.9	496.6	4.6	**
DAILY LIFE (NEVER)						
SOME LESSONS	β_{30}			-1.0	1.8	-
ABOUT HALF THE LESSONS	β_{31}			0.4	2.0	-
EVERY OR ALMOST EVERY LESSON	β_{32}			-5.8	2.2	**
Random Part						*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	213.3	346.8	2.1	6.1	0.99
Level-3: IDTEACH	v_{0kl}	4147.7	449.2	3498.4	300.5	0.16
Level-2: IDSTUD	u_{0jkl}	1428.1	36.9	1272.4	33.9	0.11
Level-1: IDCASE	e_{0ijkl}	501.8	5.8	501.8	5.8	0.00
Total Variance		6291.0		5274.7		
DIC:		172168.4		172145.1	(23.4)	
pD:		3500.7		3477.1	(23.6)	

Note: N=18625 cases, 3725 students, with 312 teachers, in 129 schools;
** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.4-7: Review Homework (Model H) compared with Model A

MCMC Estimation		*Unconditional Model (Model A)*		*Review Homework (Model H)*		
Response: BSMMATPV		Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part						
CONSTANT	β_0	477.6	3.8	492.5	4.6	**
REVIEW HOMEWORK (NEVER)						
SOME LESSONS	β_{30}			5.2	2.2	*
ABOUT HALF THE LESSONS	β_{31}			4.9	2.3	*
EVERY OR ALMOST EVERY LESSON	β_{32}			-0.1	2.3	-

Random Part					*Total Proportion Reduction*	
Level-4: IDSCHOOL	f_{0l}	164.9	249.7	39.6	71.5	0.76
Level-3: IDTEACH	v_{0kl}	4158.7	417.9	3498.7	310.2	0.16
Level-2: IDSTUD	u_{0jkl}	1429.4	37.0	1270.4	33.3	0.11
Level-1: IDCASE	e_{0ijkl}	503.3	5.8	503.3	5.9	0.00
Total Variance		6256.3		5312.0		0.15
DIC:		171899.0		171874.7	(24.3)	
pD:		3492.7		3468.4	(24.3)	

Note: N=18590 cases, 3718 students, with 312 teachers, in 129 schools;
** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.4-8: Quiz or Test (Model I) compared with Model A

MCMC Estimation		*Unconditional Model (Model A)*		*Quiz or Test (Model I)*		
Response: BSMMATPV		Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part						
CONSTANT	β_0	477.8	3.9	503.9	6.5	**
QUIZ OR TEST (NEVER)						
SOME LESSONS	β_{30}			-7.0	4.8	-
ABOUT HALF THE LESSONS	β_{31}			-14.9	5.1	**
EVERY OR ALMOST EVERY LESSON	β_{32}			-20.6	5.3	**

Random Part					*Total Proportion Reduction*	
Level-4: IDSCHOOL	f_{0l}	97.1	205.1	83.0	157.4	0.14
Level-3: IDTEACH	v_{0kl}	4263.6	410.8	3450.7	327.7	0.19
Level-2: IDSTUD	u_{0jkl}	1417.7	36.6	1249.2	32.7	0.12
Level-1: IDCASE	e_{0ijkl}	503.9	5.8	503.9	5.8	0.00
Total Variance		6282.3		5286.8		0.16
DIC:		172660.0		172633.8	(26.1)	
pD:		3506.0		3479.6	(26.4)	

Note: N=18670 cases, 3734 students, with 312 teachers, in 129 schools;
** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.4-9: Lecture-style presentation (Model J) compared with Model A

MCMC Estimation		*Unconditional Model (Model A)*		*Lecture-style presentation (Model J)*		
Response: BSMMATPV		Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part						
CONSTANT	β_0	477.1	4.1	483.9	5.8	**
LECTURE-STYLE PRESENTATION (NEVER)						
SOME LESSONS	β_{30}			7.1	4.3	-
ABOUT HALF THE LESSONS	β_{31}			13.9	4.3	**
EVERY OR ALMOST EVERY LESSON	β_{32}			12.7	4.1	**
Random Part						*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	225.3	333.8	40.3	118.5	0.82
Level-3: IDTEACH	v_{0kl}	4148.3	442.8	3489.5	318.2	0.16
Level-2: IDSTUD	u_{0jkl}	1420.9	36.9	1263.2	33.0	0.11
Level-1: IDCASE	e_{0ijkl}	502.3	5.8	502.3	5.8	0.00
Total Variance		6296.8		5295.3		0.16
DIC:		172878.8		172855.1	(23.7)	
pD:		3513.1		3489.0	(24.1)	

Note: N=18700 cases, 3740 students, with 312 teachers, in 129 schools;
** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.4-10: Calculator (Model K) compared with Model A

MCMC Estimation		*Unconditional Model (Model A)*		*Calculator (Model K)*		
Response: BSMMATPV		Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part						
CONSTANT	β_0	477.3	3.9	485.6	5.7	**
CALCULATOR (NEVER)						
SOME LESSONS	β_{30}			8.6	4.1	*
ABOUT HALF THE LESSONS	β_{31}			12.6	4.2	**
EVERY OR ALMOST EVERY LESSON	β_{32}			9.4	4.3	*
Random Part						*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	94.0	184.1	24.8	70.4	0.74
Level-3: IDTEACH	v_{0kl}	4255.8	401.3	3471.5	308.8	0.18
Level-2: IDSTUD	u_{0jkl}	1420.4	36.9	1265.3	33.2	0.11
Level-1: IDCASE	e_{0ijkl}	503.6	5.8	503.6	5.8	0.00
Total Variance		6273.8		5265.3		0.16
DIC:		172836.1		172811.8	(24.3)	
pD:		3510.7		3486.6	(24.2)	

Note: N=18690 cases, 3738 students, with 312 teachers, in 129 schools;
** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.4-11: Computer (Model L) compared with Model A

MCMC Estimation		*Unconditional Model (Model A)*		*Computer (Model L)*		
Response: BSMMATPV		Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part						
CONSTANT	β_0	477.3	4.0	497.5	4.2	**
COMPUTER (NEVER)						
SOME LESSONS	β_{30}			-2.6	1.8	-
ABOUT HALF THE LESSONS	β_{31}			-13.3	3.5	**
EVERY OR ALMOST EVERY LESSON	β_{32}			-17.3	4.0	**
Random Part						*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	216.4	287.1	49.0	97.8	0.77
Level-3: IDTEACH	v_{0kl}	4153.5	428.8	3415.0	310.1	0.18
Level-2: IDSTUD	u_{0jkl}	1420.1	36.9	1260.6	32.9	0.11
Level-1: IDCASE	e_{0ijkl}	503.5	5.8	503.5	5.8	0.00
Total Variance		6293.5		5228.1		0.17
DIC:		172598.6		172573.4	(25.1)	
pD:		3506.3		3481.4	(24.9)	

Note: N=18665 cases, 3733 students, with 312 teachers, in 129 schools;

** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.4-12: Computer for School Work (Model M) compared with Model A

MCMC Estimation		*Unconditional Model (Model A)*		*Computer for School Work (Model M)*		
Response: BSMMATPV		Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part						
CONSTANT	β_0	477.8	3.7	496.7	4.1	**
COMPUTER FOR SCHOOL WORK (NO TIME)						
A FEW TIMES A YEAR	β_{30}			3.8	1.7	*
ONCE OR TWICE A MONTH	β_{31}			-0.5	1.9	-
AT LEAST ONCE A WEEK	β_{32}			-4.1	2.6	-
EVERY DAY	β_{33}			-30.5	4.9	**
Random Part						*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	222.9	321.2	80.6	142.9	0.64
Level-3: IDTEACH	v_{0kl}	4111.6	440.1	3311.9	315.7	0.19
Level-2: IDSTUD	u_{0jkl}	1431.2	37.1	1262.2	33.3	0.12
Level-1: IDCASE	e_{0ijkl}	501.9	5.8	501.9	5.9	0.00
Total Variance		6267.6		5156.6		0.18
DIC:		171619.0		171591.3	(27.6)	
pD:		3489.9		3463.2	(26.6)	

Note: N=18565 cases, 3713 students, with 312 teachers, in 129 schools;

** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.4-13: Procedures for Complex Problems (Model N) compared with Model A

MCMC Estimation		*Unconditional Model (Model A)*		*Complex Problems (Model N)*		
Response: BSMMATPV		Coefficient	s.e.	Coef.	s.e.	Sig.
Fixed Part						
CONSTANT	β_0	477.6	4.1	494.8	4.5	**
PROCEDURES FOR COMPLEX PROB (NEVER)						
SOME LESSONS	β_{30}			4.1	1.9	*
ABOUT HALF THE LESSONS	β_{31}			1.3	2.1	-
EVERY OR ALMOST EVERY LESSON	β_{32}			0.8	2.4	-
Random Part						*Total Proportion Reduction*
Level-4: IDSCHOOL	f_{0l}	177.4	283.1	37.1	76.0	0.79
Level-3: IDTEACH	v_{0kl}	4174.8	432.0	3495.2	312.2	0.16
Level-2: IDSTUD	u_{0jkl}	1430.1	37.2	1277.2	33.9	0.11
Level-1: IDCASE	e_{0ijkl}	500.3	5.8	500.4	5.8	0.00
Total Variance		6282.7		5309.8		0.15
DIC:		170821.8		170800.9	(20.9)	
pD:		3475.2		3453.5	(21.7)	

Note: N=18485 cases, 3697 students, with 312 teachers, in 129 schools;

** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Appendix 5.4.2. Model of mathematics achievement (G8) with ALL control and predictor variables

Table 5.4-14: Modelling all predictor and control variables (G8 Mathematics - Controls)

BSMMATPV	BMATH_ALL_COMBINED_MODEL_IGLS				
Response	Coef.	S.E.	Sig	effect size	+/- error
OWN LANGUAGE					
ALMOST ALWAYS	-1.8	3.0	-	-0.05	-0.08
SOMETIMES	3.9	4.3	-	0.10	0.11
NEVER	-14.5	7.4	-	-0.38	-0.20
BOOKS AT HOME					
NONE OR VERY FEW (0 TO 10 BOOKS)	-15.0	2.1	**	-0.40	-0.06
ONE SHELF (11 TO 25 BOOKS)	-8.2	1.9	**	-0.22	-0.05
TWO BOOKCASES (101 TO 200 BOOKS)	2.8	2.1	-	0.07	0.06
THREE OR MORE BOOKCASES (OVER 200)	13.9	2.2	**	0.37	0.06
FORMATIVE YEARS IN COUNTRY(UK)					
ARRIVED IN UK AGED 5 TO 10 YEARS	4.5	7.1	-	0.12	0.19
ARRIVED IN UK OLDER THAN 10 YEARS	-14.7	5.4	**	-0.39	-0.14
GENDER					
GIRL	-11.3	1.5	**	-0.33	-0.04
WORK IN PAID JOB					
WKPJ_LESS THAN 1 HOUR	-4.9	2.5	-	-0.13	-0.07
WKPJ_1 TO 2 HOURS	-1.3	2.6	-	-0.03	-0.07
WKPJ_MORE THAN 2 BUT LESS THAN 4 HRS	-7.3	3.3	*	-0.19	-0.09
WKPJ_4 OR MORE HOURS	-12.0	3.3	**	-0.32	-0.09
PLAY SPORTS					
PLSP_LESS THAN 1 HOUR	0.2	2.3	-	0.01	0.06
PLSP_1 TO 2 HOURS	4.5	2.2	*	0.12	0.06
PLSP_MORE THAN 2 BUT LESS THAN 4 HOURS	-0.9	2.4	-	-0.02	-0.06
PLSP_4 OR MORE HOURS	-4.2	2.5	-	-0.11	-0.07
OUT-OF-SCHOOL ACTIVITIES					
(BSCLSOC-gm)^1	-6.7	1.3	**	-0.30	-0.04
(BSCLSOC-gm)^2	-2.7	1.3	**	-0.35	-0.07
(BSCLIND-gm)^1	0.2	1.1	*	-0.14	-0.07
(BSCLIND-gm)^2	-2.2	1.1	-	0.02	0.06
(BSCLPD-gm)^1	2.0	2.1	*	-0.12	-0.06
(BSCLPD-gm)^2	-11.8	3.3	-	0.10	0.11
(BSCLPD-gm)^3	5.5	2.0	**	-0.62	-0.17

Note: ** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

Table 5.4-15: Modelling all predictor and control variables (G8 Mathematics - Predictors)

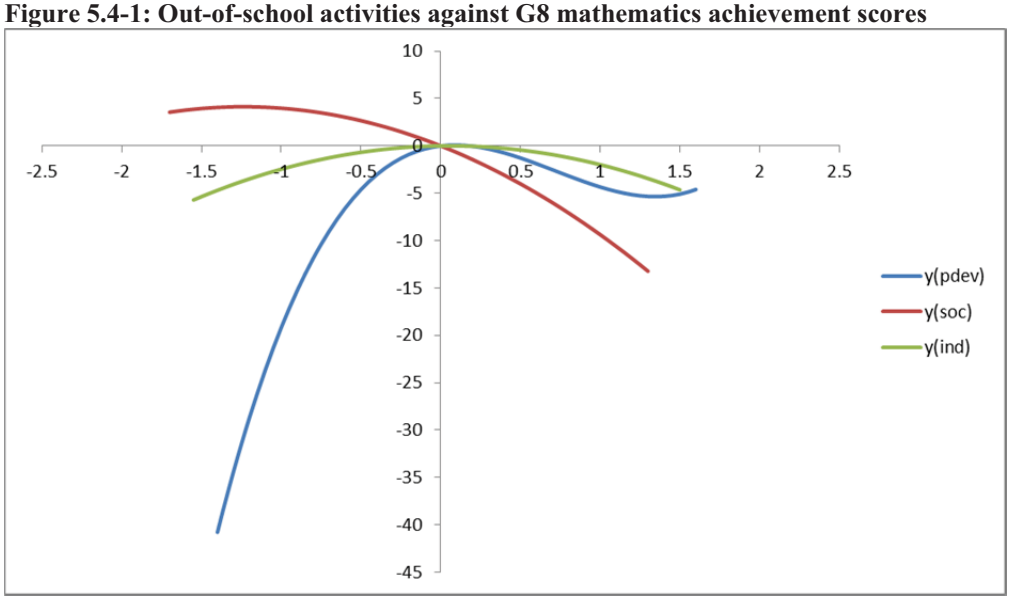
BSMMATPV	G8MATH_ALL_COMBINED_MODEL_IGLS				
Response	Coef.	S.E.	Sig	effect size	+/- error
BS4MHFRR_R:SOME LESSONS	10.6	2.2	**	0.28	0.06
BS4MHFRR_R:ABOUT HALF THE LESSONS	13.4	2.4	**	0.36	0.06
BS4MHFRR_R:EVERY OR ALMOST EVERY LES.	17.2	2.6	**	0.46	0.07
BS4MHEXP_R:SOME LESSONS	2.9	3.5	-	0.08	0.09
BS4MHEXP_R:ABOUT HALF THE LESSONS	4.7	3.5	-	0.13	0.09
BS4MHEXP_R:EVERY OR ALMOST EVERY LES.	4.6	3.4	-	0.12	0.09
BS4MHMDL_R:SOME LESSONS	-2.6	1.9	-	-0.07	-0.05
BS4MHMDL_R:ABOUT HALF THE LESSONS	-1.4	2.2	-	-0.04	-0.06
BS4MHMDL_R:EVERY OR ALMOST EVERY LES.	-5.2	2.5	*	-0.14	-0.07
BS4MHROH_R:SOME LESSONS	4.7	2.3	*	0.12	0.06
BS4MHROH_R:ABOUT HALF THE LESSONS	3.3	2.4	-	0.09	0.06
BS4MHROH_R:EVERY OR ALMOST EVERY LES.	0.1	2.5	-	0.00	0.07
BS4MHHQT_R:SOME LESSONS	-15.0	5.1	**	-0.40	-0.14
BS4MHHQT_R:ABOUT HALF THE LESSONS	-21.3	5.4	**	-0.57	-0.14
BS4MHHQT_R:EVERY OR ALMOST EVERY LES.	-22.4	5.7	**	-0.60	-0.15
BS4MHWSG_R:SOME LESSONS	-1.0	1.7	-	-0.03	-0.04
BS4MHWSG_R:ABOUT HALF THE LESSONS	-1.9	2.4	-	-0.05	-0.06
BS4MHWSG_R:EVERY OR ALMOST EVERY LES.	-13.0	3.1	**	-0.35	-0.08
BS4MHWPO_R:SOME LESSONS	3.7	3.5	-	0.10	0.09
BS4MHWPO_R:ABOUT HALF THE LESSONS	6.2	3.4	-	0.16	0.09
BS4MHWPO_R:EVERY OR ALMOST EVERY LES.	14.2	3.5	**	0.38	0.09
BS4MHLSP_R:SOME LESSONS	3.1	4.6	-	0.08	0.12
BS4MHLSP_R:ABOUT HALF THE LESSONS	8.9	4.6	-	0.24	0.12
BS4MHLSP_R:EVERY OR ALMOST EVERY LES.	6.3	4.5	-	0.17	0.12
BS4MHCAL_R:SOME LESSONS	6.4	4.4	-	0.17	0.12
BS4MHCAL_R:ABOUT HALF THE LESSONS	11.3	4.6	*	0.30	0.12
BS4MHCAL_R:EVERY OR ALMOST EVERY LES.	11.6	4.7	*	0.31	0.13
BS4MHCOM_R:SOME LESSONS	-2.4	1.9	-	-0.06	-0.05
BS4MHCOM_R:ABOUT HALF THE LESSONS	-8.8	3.7	*	-0.23	-0.10
BS4MHCOM_R:EVERY OR ALMOST EVERY LES.	-6.9	4.8	-	-0.18	-0.13
BS4MCSWM_R:A FEW TIMES A YEAR	4.8	1.8	**	0.13	0.05
BS4MCSWM_R:ONCE OR TWICE A MONTH	1.0	1.9	-	0.03	0.05
BS4MCSWM_R:AT LEAST ONCE A WEEK	-1.2	2.7	-	-0.03	-0.07
BS4MCSWM_R:EVERY DAY	-24.2	5.3	**	-0.64	-0.14

BSMMATPV	G8MATH_ALL_COMBINED_MODEL_IGLS				
Response	Coef.	S.E.	Sig	effect size	+/- error
BS4MHGCT_R:SOME LESSONS	1.7	3.5	-	0.05	0.09
BS4MHGCT_R:ABOUT HALF THE LESSONS	-6.5	3.8	-	-0.17	-0.10
BS4MHGCT_R:EVERY OR ALMOST EVERY LES.	-13.1	4.3	**	-0.35	-0.11
BS4MHSCP_R:SOME LESSONS	3.5	2.0	-	0.09	0.05
BS4MHSCP_R:ABOUT HALF THE LESSONS	1.3	2.3	-	0.04	0.06
BS4MHSCP_R:EVERY OR ALMOST EVERY LES.	3.6	2.8	-	0.10	0.07

Note: ** t-value>2.58 a CI of 99%; * t-value>1.96 a CI of 95%

The best fitting polynomials for the out-of-school interests in the G8 mathematics achievement data were a cubic for the Personal Development dimension and quadratic models for each of Individual and Social dimensions. There are moderate to high effect sizes attributed to the component parts of the Personal Development dimension, but smaller effects reported on the Social dimension. The Individual dimension offers a much weaker association with mathematics achievement scores, only significant at the 5 per cent level and with small effect sizes. Essentially the cubic model for personal development (do jobs at home; read book for enjoyment; do homework) appears to have the strongest effect in that low levels of engagement are associated with very low negative coefficients, and high levels of engagement dip slightly, possibly reflecting on those who spend a lot of time studying because of their particular educational needs.

Figure 5.4-1: Out-of-school activities against G8 mathematics achievement scores



The function $y = 5.46x^3 - 11.76x^2 + 1.96x$ is used to model the significant residuals for personal development dimension. The flattened curve for the Individual

dimension, providing only a small effect in the model, is represented by the function $y = -2.22x^2 + 0.24x$. These graphical representations reflect the effect sizes reported in Table 5.4-14 and support interpretation of the overall findings. As with all the previous graphs the domain for each function has been restricted to suit the data, avoiding any misrepresentation by extending the graph beyond observed values.

Appendix 6. Alternative Analyses

Appendix 6.1. Grade 4

Appendix 6.1.1. Alternative Models for Mathematics and Science – MLPV, AVPV, SPV, Rubin

Table 6.1-1: G4 MATHS COMPARING ALTERNATIVE MODELS

	MLPV (ASMMATPV)		AVPV (ASMMAT_AV)		SPV (ASMMATO3)		ASMMAT_RUBIN	
	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size
EXPLAIN								
SOME LESSONS	9.2 (3.3)**	0.15	9.2 (3.3)**	0.14	7.5 (3.6)*	0.11	9.2 (3.8)*	0.13
ABOUT HALF THE LESSONS	13.4 (3.6)**	0.21	13.4 (3.6)**	0.21	12.5 (3.9)**	0.18	13.5 (4.0)**	0.20
EVERY OR ALMOST EVERY LESS.	11.1 (3.7)**	0.18	11.1 (3.7)**	0.17	10.2 (4.1)*	0.15	11.2 (4.2)**	0.16
MEMORISE								
SOME LESSONS	3.4 (4.4)-	0.05	3.4 (4.4)-	0.05	5.5 (4.8)-	0.08	3.4 (5.3)-	0.05
ABOUT HALF THE LESSONS	7.2 (4.4)-	0.11	7.2 (4.4)-	0.11	8.1 (4.8)-	0.12	7.2 (5.3)-	0.10
EVERY OR ALMOST EVERY LESS.	18.8 (4.4)**	0.30	18.8 (4.4)**	0.29	18.9 (4.9)**	0.28	18.7 (5.1)**	0.27
MEASURE								
SOME LESSONS	1.4 (3.0)-	0.02	1.4 (3.0)-	0.02	1.7 (3.3)-	0.03	1.3 (3.5)-	0.02
ABOUT HALF THE LESSONS	-16.9 (4.2)**	-0.27	-16.9 (4.2)**	-0.26	-15.9 (4.6)**	-0.23	-16.8 (5.9)**	-0.25
EVERY OR ALMOST EVERY LESS.	-28.7 (4.8)**	-0.45	-28.7 (4.8)**	-0.44	-24.6 (5.2)**	-0.36	-28.7 (6.1)**	-0.42
TABLES, CHARTS & GRAPHS								
SOME LESSONS	25.0 (4.0)**	0.39	25.0 (4.0)**	0.39	22.9 (4.4)**	0.33	25.1 (4.7)**	0.37
ABOUT HALF THE LESSONS	20.5 (4.3)**	0.32	20.5 (4.3)**	0.32	19.7 (4.7)**	0.29	20.6 (4.9)**	0.30
EVERY OR ALMOST EVERY LESS.	14.1 (4.8)**	0.22	14.1 (4.8)**	0.22	10.5 (5.2)*	0.15	14.1 (6.2)*	0.21

Table 6.1-1: G4 MATHS COMPARING ALTERNATIVE MODELS (Continued)

	MLPV (ASMMATPV)		AVPV (ASMMAT_AV)		SPV (ASMMATO3)		ASMMAT_RUBIN	
	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size
WORK ON OWN								
SOME LESSONS	22.8 (7.7)**	0.36	22.8 (7.7)**	0.35	27.0 (8.4)**	0.39	23.3 (8.9)**	0.34
ABOUT HALF THE LESSONS	28.9 (7.7)**	0.46	28.9 (7.7)**	0.45	29.9 (8.4)**	0.43	29.3 (8.4)**	0.43
EVERY OR ALMOST EVERY LESSON	40.5 (7.5)**	0.64	40.5 (7.5)**	0.63	42.8 (8.3)**	0.62	41.0 (8.4)**	0.60
WORK IN SMALL GROUP								
SOME LESSONS	11.5 (4.1)**	0.18	11.5 (4.1)**	0.18	13.5 (4.5)**	0.20	11.5 (5.2)*	0.17
ABOUT HALF THE LESSONS	15.5 (4.5)**	0.24	15.5 (4.5)**	0.24	17.3 (4.9)**	0.25	15.5 (5.7)**	0.23
EVERY OR ALMOST EVERY LESSON	-1.0 (4.5)-	-0.02	-1.0 (4.5)-	-0.02	0.1 (4.9)-	0.00	-1.2 (5.9)-	-0.02
USE A CALCULATOR IN CLASS								
SOME LESSONS	12.4 (2.3)**	0.20	12.4 (2.3)**	0.19	12.9 (2.5)**	0.19	12.6 (2.6)**	0.18
ABOUT HALF THE LESSONS	4.5 (4.6)-	0.07	4.5 (4.6)-	0.07	7.7 (5.0)-	0.11	4.5 (5.5)-	0.07
EVERY OR ALMOST EVERY LESSON	-21.8 (7.5)**	-0.34	-21.8 (7.5)**	-0.34	-20.6 (8.2)*	-0.30	-21.8 (9.6)*	-0.32
USE A COMPUTER IN CLASS								
SOME LESSONS	-3.2 (2.3)-	-0.05	-3.2 (2.3)-	-0.05	-5.6 (2.5)*	-0.08	-3.3 (3.2)-	-0.05
ABOUT HALF THE LESSONS	-13.1 (3.7)**	-0.21	-13.1 (3.7)**	-0.20	-13.7 (4.0)**	-0.20	-13.0 (4.1)**	-0.19
EVERY OR ALMOST EVERY LESSON	-30.3 (4.4)**	-0.48	-30.3 (4.4)**	-0.47	-36.2 (4.8)**	-0.53	-30.3 (7.0)**	-0.44

Note: ** t-value>2.58, sig. level 1%; * t-value>1.96, sig. level 5%

Table 6.1-2: G4 SCIENCE COMPARING ALTERNATIVE MODELS

	MLPV (ASSSCIPV)		AVPV (ASSSCI_AV)		SPV (ASSSCIO3)		ASSSCI_RUBIN	
Response	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size
EXPLAIN SCIENCE STUDIED (NEVER)								
A FEW TIMES A YEAR	-1.3 (3.2)-	-0.02	-1.3 (3.2)-	-0.02	-2.3 (3.7)-	-0.03	-1.2 (5.1)-	-0.02
ONCE OR TWICE A MONTH	0.0 (3.3)-	0.00	0.0 (3.3)-	0.00	-3.0 (3.8)-	-0.04	0.1 (5.0)-	0.00
AT LEAST ONCE A WEEK	-2.6 (3.6)-	-0.04	-2.6 (3.6)-	-0.04	-4.9 (4.1)-	-0.07	-2.4 (5.0)-	-0.04
MEMORISE SCIENCE FACTS (NEVER)								
A FEW TIMES A YEAR	4.5 (3.2)-	0.07	4.5 (3.2)-	0.07	3.4 (3.6)-	0.05	4.4 (4.0)-	0.06
ONCE OR TWICE A MONTH	13.6 (3.2)**	0.22	13.6 (3.2)**	0.22	11.7 (3.6)**	0.17	13.4 (4.7)**	0.20
AT LEAST ONCE A WEEK	24.7 (3.2)**	0.41	24.7 (3.2)**	0.40	23.5 (3.7)**	0.34	24.6 (4.0)**	0.36
WATCH SCIENCE EXPERIMENTS (NEVER)								
A FEW TIMES A YEAR	14.3 (3.3)**	0.24	14.3 (3.3)**	0.23	14.8 (3.7)**	0.22	14.3 (4.5)**	0.21
ONCE OR TWICE A MONTH	3.3 (3.4)-	0.06	3.3 (3.4)-	0.05	4.3 (3.9)-	0.06	3.5 (4.4)-	0.05
AT LEAST ONCE A WEEK	-8.0 (3.6)*	-0.13	-8.0 (3.6)*	-0.13	-8.3 (4.1)*	-0.12	-7.8 (4.9)-	-0.11
PLAN SCIENCE EXPERIMENTS (NEVER)								
A FEW TIMES A YEAR	6.3 (2.7)*	0.10	6.3 (2.7)*	0.10	4.7 (3.0)-	0.07	6.3 (3.5)-	0.09
ONCE OR TWICE A MONTH	0.6 (3.0)-	0.01	0.6 (3.0)-	0.01	2.3 (3.4)-	0.03	0.6 (4.3)-	0.01
AT LEAST ONCE A WEEK	-0.7 (3.7)-	-0.01	-0.7 (3.7)-	-0.01	-3.9 (4.2)-	-0.06	-0.6 (5.5)-	-0.01
DO SCIENCE EXPERIMENTS (NEVER)								
A FEW TIMES A YEAR	12.5 (3.1)**	0.21	12.5 (3.1)**	0.20	12.5 (3.5)**	0.18	12.6 (4.9)**	0.19
ONCE OR TWICE A MONTH	15.6 (3.4)**	0.26	15.6 (3.4)**	0.25	17.4 (3.8)**	0.25	15.7 (4.9)**	0.23
AT LEAST ONCE A WEEK	10.7 (3.8)**	0.18	10.7 (3.8)**	0.17	12.4 (4.4)**	0.18	10.7 (5.3)*	0.16

Table 6.1-2: G4 SCIENCE COMPARING ALTERNATIVE MODELS (Continued)

Response	MLPV (ASSSCIPV) COEF (S.E.)Sig	effect size	AVPV (ASSSCI_AV) COEF (S.E.)Sig	effect size	SPV (ASSSCIO3) COEF (S.E.)Sig	effect size	ASSSCI_RUBIN COEF (S.E.)Sig	effect size
WORK ON OWN(NEVER)								
A FEW TIMES A YEAR	9.2 (4.0)*	0.15	9.2 (4.0)*	0.15	7.1 (4.5)-	0.10	9.1 (4.8)-	0.13
ONCE OR TWICE A MONTH	12.6 (3.7)**	0.21	12.6 (3.7)**	0.20	11.6 (4.2)**	0.17	12.6 (4.7)**	0.19
AT LEAST ONCE A WEEK	15.1 (3.6)**	0.25	15.1 (3.6)**	0.24	15.0 (4.1)**	0.22	15.1 (4.8)**	0.22
WORK IN GROUP (NEVER)								
A FEW TIMES A YEAR	2.8 (3.8)-	0.05	2.8 (3.8)-	0.05	1.7 (4.3)-	0.02	2.9 (7.5)-	0.00
ONCE OR TWICE A MONTH	0.2 (3.8)-	0.00	0.2 (3.8)-	0.00	-1.4 (4.3)-	-0.02	0.5 (6.4)-	0.04
AT LEAST ONCE A WEEK	-5.3 (4.0)-	-0.09	-5.3 (4.0)-	-0.09	-6.0 (4.5)-	-0.09	-5.0 (7.3)-	0.01
COMPUTER USE (NEVER)								
SOME LESSONS	0.5 (2.6)-	0.01	0.5 (2.6)-	0.01	-0.4 (2.9)-	-0.01	0.3 (3.0)-	-0.07
ABOUT HALF THE LESSONS	-8.5 (2.9)**	-0.14	-8.5 (2.9)**	-0.14	-7.6 (3.2)*	-0.11	-8.7 (3.7)*	0.00
EVERY OR ALMOST EVERY LESS.	-20.2 (3.6)**	-0.33	-20.2 (3.6)**	-0.33	-21.1 (4.1)**	-0.31	-20.5 (5.2)**	0.00

Note: ** t-value>2.58, sig. level 1%; * t-value>1.96, sig. level 5%

Appendix 6.2. Grade 8

Appendix 6.2.1. Alternative Models for Mathematics and Science – MLPV, AVPV, SPV, Rubin

Table 6.2-1: G8 MATHS COMPARING ALTERNATIVE MODELS

	MLPV (BSMMATPV)		AVPV (BSMMAT_AV)		SPV (BSMMATO3)		BSMMAT_RUBIN	
	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size
MEMORISE								
SOME LESSONS	10.6 (2.2)**	0.28	10.6 (2.2)**	0.27	13.4 (2.5)**	0.30	11.0 (2.9)**	0.25
ABOUT HALF THE LESSONS	13.4(2.4)**	0.36	13.4(2.4)**	0.34	16.1 (2.8)**	0.37	14.1 (3.6)**	0.32
EVERY OR ALMOST EVERY LESS.	17.2 (2.6)**	0.46	17.2 (2.6)**	0.44	19.8 (3.0)**	0.45	17.9 (3.2)**	0.41
EXPLAIN								
SOME LESSONS	2.9 (3.5)-	0.08	2.9 (3.5)-	0.08	0.7 (4.0)-	0.02	2.6 (5.0)-	0.06
ABOUT HALF THE LESSONS	4.7 (3.5)-	0.13	4.7 (3.5)-	0.12	0.0 (4.1)-	0.00	4.4 (5.2)-	0.10
EVERY OR ALMOST EVERY LESS.	4.6 (3.4)-	0.12	4.6 (3.4)-	0.12	2.4 (3.9)-	0.06	4.3 (4.8)-	0.10
WORK IN SMALL GROUP								
SOME LESSONS	-1.0 (1.7)-	-0.03	-1.0 (1.7)-	-0.03	-2.8 (1.9)-	-0.06	-1.0 (2.3)-	-0.02
ABOUT HALF THE LESSONS	-1.9 (2.4)-	-0.05	-1.9 (2.4)-	-0.05	-3.7 (2.7)-	-0.08	-1.9 (3.1)-	-0.04
EVERY OR ALMOST EVERY LESS.	-13.0 (3.1)**	-0.35	-13.0 (3.1)**	-0.33	-14.6 (3.6)**	-0.33	-13.2 (4.2)**	-0.30
WORK ON OWN								
SOME LESSONS	3.7 (3.5)-	0.10	3.7 (3.5)-	0.10	5.6 (4.0)-	0.13	3.9 (5.0)-	0.09
ABOUT HALF THE LESSONS	6.2 (3.4)-	0.16	6.2 (3.4)-	0.16	7.3 (4.0)-	0.16	6.3 (4.3)-	0.15
EVERY OR ALMOST EVERY LESS.	14.2 (3.5)**	0.38	14.2 (3.5)**	0.36	16.0 (4.0)**	0.36	14.5 (4.6)**	0.33

Table 6.2-1: G8 MATHS COMPARING ALTERNATIVE MODELS (Continued)

	MLPV (BSMMATPV)		AVPV (BSMMAT_AV)		SPV (BSMMATO3)		BSMMAT_RUBIN	
	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size
REVIEW HOMEWORK								
SOME LESSONS	4.7 (2.3)*	0.12	4.7 (2.3)*	0.12	4.8 (2.6)-	0.11	4.9 (3.3)-	0.11
ABOUT HALF THE LESSONS	3.3 (2.4)-	0.09	3.3 (2.4)-	0.09	5.4 (2.8)-	0.12	3.6 (3.4)-	0.08
EVERY OR ALMOST EVERY LESS.	0.1 (2.5)-	0.00	0.1 (2.5)-	0.00	-0.2 (2.8)-	0.00	0.4 (3.2)-	0.01
LECTURE-STYLE PRESENTATION								
SOME LESSONS	3.1 (4.6)-	0.08	3.1 (4.6)-	0.08	-2.2 (5.3)-	-0.05	3.0 (6.4)-	0.07
ABOUT HALF THE LESSONS	8.9 (4.6)-	0.24	8.9 (4.6)-	0.23	7.1 (5.3)-	0.16	8.8 (5.6)-	0.20
EVERY OR ALMOST EVERY LESS.	6.3 (4.5)-	0.17	6.3 (4.5)-	0.16	4.9 (5.2)-	0.11	6.4 (5.3)-	0.15
USE A CALCULATOR IN CLASS								
SOME LESSONS	6.4 (4.4)-	0.17	6.4 (4.4)-	0.16	1.5 (5.1)-	0.03	6.7 (9.1)-	0.15
ABOUT HALF THE LESSONS	11.3 (4.6)*	0.30	11.3 (4.6)*	0.29	4.9 (5.3)-	0.11	12.0 (10.2)-	0.27
EVERY OR ALMOST EVERY LESS.	11.6 (4.7)*	0.31	11.6 (4.7)*	0.30	6.6 (5.4)-	0.15	12.4 (9.4)-	0.28
USE A COMPUTER IN CLASS								
SOME LESSONS	-2.4 (1.9)-	-0.06	-2.4 (1.9)-	-0.06	-4.0 (2.2)-	-0.09	-2.8 (2.9)-	-0.06
ABOUT HALF THE LESSONS	-8.8 (3.7)*	-0.23	-8.8 (3.7)*	-0.22	-6.5 (4.3)-	-0.15	-9.4 (5.5)-	-0.21
EVERY OR ALMOST EVERY LESS.	-6.9 (4.8)-	-0.18	-6.9 (4.8)-	-0.18	-12.8 (5.5)*	-0.29	-7.6 (6.6)-	-0.17
COMPUTER FOR SCH WORK								
A FEW TIMES A YEAR	4.8 (1.8)**	0.13	4.8 (1.8)**	0.12	3.6 (2.0)-	0.08	4.9 (3.2)-	0.11
ONCE OR TWICE A MONTH	1.0 (1.9)-	0.03	1.0 (1.9)-	0.03	1.8 (2.2)-	0.04	1.0 (3.5)-	0.02
AT LEAST ONCE A WEEK	-1.2 (2.7)-	-0.03	-1.2 (2.7)-	-0.03	-1.4 (3.1)-	-0.03	-1.6 (3.7)-	-0.04
EVERY DAY	-24.2 (5.3)**	-0.64	-24.2 (5.3)**	-0.62	-22.9 (6.1)**	-0.52	-25.0 (8.2)**	-0.57

Note: ** t-value>2.58, sig. level 1%; * t-value>1.96, sig. level 5%

Table 6.2-1: G8 MATHS COMPARING ALTERNATIVE MODELS (Continued)

	MLPV (BSMMATPV)		AVPV (BSMMAT_AV)		SPV (BSMMATO3)		BSMMAT_RUBIN	
	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size
INTERPRET GRAPHS CHARTS TABLES								
SOME LESSONS	1.7 (3.5)-	0.05	1.7 (3.5)-	0.04	1.7 (4.1)-	0.04	2.0 (4.4)-	0.05
ABOUT HALF THE LESSONS	-6.5 (3.8)-	-0.17	-6.5 (3.8)-	-0.17	-7.4 (4.4)-	-0.17	-6.4 (4.7)-	-0.15
EVERY OR ALMOST EVERY LESS.	-13.1 (4.3)**	-0.35	-13.1 (4.3)**	-0.34	-13.5 (5.0)**	-0.31	-13.2 (5.6)*	-0.30
SOLVE COMPLEX PROBLEMS								
SOME LESSONS	3.5 (2.0)-	0.09	3.5 (2.0)-	0.09	1.9 (2.3)-	0.04	3.6 (3.9)-	0.08
ABOUT HALF THE LESSONS	1.3 (2.3)-	0.04	1.3 (2.3)-	0.03	-0.5 (2.6)-	-0.01	1.4 (3.9)-	0.03
EVERY OR ALMOST EVERY LESS.	3.6 (2.8)-	0.09	3.6 (2.8)-	0.09	1.9 (3.2)-	0.04	3.6 (4.3)-	0.08
DAILY LIFE								
SOME LESSONS	-2.6 (1.9)-	-0.07	-2.6 (1.9)-	-0.07	-2.1 (2.2)-	-0.05	-2.7 (2.7)-	-0.06
ABOUT HALF THE LESSONS	-1.4 (2.2)-	-0.04	-1.4 (2.2)-	-0.04	1.1 (2.5)-	0.03	-1.7 (3.2)-	-0.04
EVERY OR ALMOST EVERY LESS.	-5.2 (2.5)*	-0.14	-5.2 (2.5)*	-0.13	-5.4 (2.8)-	-0.12	-5.7 (2.9)-	-0.13
QUIZ OR TEST								
SOME LESSONS	-15.0 (5.1)**	-0.40	-15.0 (5.1)**	-0.39	-15.5 (5.9)**	-0.35	-14.8 (6.5)*	-0.34
ABOUT HALF THE LESSONS	-21.3 (5.4)**	-0.57	-21.3 (5.4)**	-0.55	-22.6 (6.2)**	-0.51	-21.1 (6.9)**	-0.48
EVERY OR ALMOST EVERY LESS.	-22.4 (5.7)**	-0.60	-22.4 (5.7)**	-0.58	-25.4 (6.5)**	-0.58	-22.5 (6.9)**	-0.51

Table 6.2-2: G8 SCIENCE COMPARING ALTERNATIVE MODELS

	MLPV (BSSSCIPV)		AVPV (BSSSCI_AV)		SPV (BSSSCIO3)		BSSSCI_RUBIN	
	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size
MEMORISE SCI FACTS ETC.								
SOME LESSONS	3.1 (4.0)-	0.05	3.1 (4.0)-	0.05	2.0 (4.4)-	0.03	3.3 (5.7)-	0.05
ABOUT HALF THE LESS.	7.3 (4.2)-	0.13	7.3 (4.2)-	0.13	5.9 (4.6)-	0.10	7.6 (5.1)-	0.12
EVERY OR ALMOST EVERY L.	7.0 (4.4)-	0.12	7.0 (4.4)-	0.12	6.7 (4.8)-	0.11	7.3 (6.0)-	0.12
EXPLAIN SCIENCE STUDIED								
SOME LESSONS	3.2 (4.3)-	0.06	3.2 (4.3)-	0.05	0.9 (4.8)-	0.01	3.2 (5.4)-	0.05
ABOUT HALF THE LESSONS	8.2 (4.5)-	0.15	8.2 (4.5)-	0.14	5.8 (4.9)-	0.09	8.3 (5.9)-	0.13
EVERY OR ALMOST EVERY L.	12.7 (4.6)**	0.22	12.7 (4.6)**	0.22	10.1 (5.1)*	0.16	12.8 (5.9)*	0.20
CONDUCT EXPERIMENT								
SOME LESSONS	5.5 (8.0)-	0.10	5.5 (8.0)-	0.10	-3.0 (8.8)-	-0.05	5.3 (10.6)-	0.09
ABOUT HALF THE LESSONS	9.0 (8.1)-	0.16	9.0 (8.1)-	0.16	0.8 (8.9)-	0.01	8.9 (10.4)-	0.14
EVERY OR ALMOST EVERY L.	8.9 (8.2)-	0.16	8.9 (8.2)-	0.15	2.7 (9.1)-	0.04	8.7 (10.0)-	0.14
WATCH TEACHER DEMO.								
SOME LESSONS	-5.9 (7.4)-	-0.11	-5.9 (7.4)-	-0.10	-13.4 (8.2)-	-0.22	-6.2 (9.8)-	-0.10
ABOUT HALF THE LESS.	0.2 (7.4)-	0.00	0.2 (7.4)-	0.00	-6.6 (8.2)-	-0.11	0.1 (9.3)-	0.00
EVERY OR ALMOST EVERY L.	-6.1 (7.5)-	-0.11	-6.1 (7.5)-	-0.11	-14.5 (8.2)-	-0.23	-6.3 (10.0)-	-0.10
PLAN EXPERIMENT OR INVEST.								
SOME LESSONS	1.7 (4.3)-	0.03	1.7 (4.3)-	0.03	-2.1 (4.7)-	-0.03	2.0 (6.3)-	0.03
ABOUT HALF THE LESS.	0.9 (4.4)-	0.02	0.9 (4.4)-	0.02	0.5 (4.9)-	0.01	1.2 (5.6)-	0.02
EVERY OR ALMOST EVERY L.	-3.4 (4.7)-	-0.06	-3.4 (4.7)-	-0.06	-6.1 (5.2)-	-0.10	-3.1 (5.7)-	-0.05

Table 6.2-2: G8 SCIENCE COMPARING ALTERNATIVE MODELS (Continued)

	MLPV (BSSSCIPV)		AVPV (BSSSCI_AV)		SPV (BSSSCIO3)		BSSSCI_RUBIN	
	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size
WORK ON OWN								
SOME LESSONS	2.6 (3.7)-	0.05	2.6 (3.7)-	0.04	2.0 (4.0)-	0.03	2.6 (4.8)-	0.04
ABOUT HALF THE LESSONS	6.9 (3.7)-	0.12	6.9 (3.7)-	0.12	5.9 (4.1)-	0.09	6.9 (4.5)-	0.11
EVERY OR ALMOST EVERY L.	10.9 (4.0)**	0.19	10.9 (4.0)**	0.19	9.2 (4.4)*	0.15	10.8 (5.2)*	0.17
WORK IN GROUP								
SOME LESSONS	4.5 (9.3)-	0.08	4.5 (9.3)-	0.08	6.9 (10.2)-	0.11	4.6 (16.0)-	0.07
ABOUT HALF THE LESSONS	11.0 (9.3)-	0.19	11.0 (9.3)-	0.19	12.8 (10.2)-	0.21	11.2 (15.0)-	0.18
EVERY OR ALMOST EVERY L.	9.5 (9.3)-	0.17	9.5 (9.3)-	0.16	10.1 (10.2)-	0.16	9.6 (14.3)-	0.15
REVIEW HOMEWORK								
SOME LESSONS	-5.0 (2.6)-	-0.09	-5.0 (2.6)-	-0.09	-6.1 (2.8)*	-0.10	-5.0 (3.5)-	-0.08
ABOUT HALF THE LESSONS	-12.7 (2.9)**	-0.23	-12.7 (2.9)**	-0.22	-13.1 (3.2)**	-0.21	-12.8 (3.4)**	-0.20
EVERY OR ALMOST EVERY L	-15.6 (3.1)**	-0.28	-15.6 (3.1)**	-0.27	-16.5 (3.4)**	-0.27	-15.7 (3.9)**	-0.25
QUIZ OR TEST								
SOME LESSONS	7.5 (6.0)-	0.13	7.5 (6.0)-	0.13	8.4 (6.6)-	0.14	7.7 (9.9)-	0.12
ABOUT HALF THE LESSONS	-0.5 (6.4)-	-0.01	-0.5 (6.4)-	-0.01	-0.8 (7.0)-	-0.01	-0.4 (8.9)-	-0.01
EVERY OR ALMOST EVERY L	-14.2 (6.8)*	-0.25	-14.2 (6.8)*	-0.25	-12.7 (7.4)-	-0.20	-14.3 (12.1)-	-0.23
LECTURE-STYLE PRESENT.								
SOME LESSONS	1.1 (6.9)-	0.02	1.1 (6.9)-	0.02	1.2 (7.6)-	0.02	0.9 (8.7)-	0.01
ABOUT HALF THE LESSONS	6.9 (6.8)-	0.12	6.9 (6.8)-	0.12	9.3 (7.5)-	0.15	6.7 (9.2)-	0.11
EVERY OR ALMOST EVERY L	14.3 (6.7)*	0.25	14.3 (6.7)*	0.25	17.1 (7.4)*	0.28	14.4 (8.9)-	0.23

Table 6.2-2: G8 SCIENCE COMPARING ALTERNATIVE MODELS (Continued)

	MLPV (BSSSCIPV)		AVPV (BSSSCI_AV)		SPV (BSSSCIO3)		BSSSCI_RUBIN	
	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size	COEF (S.E.)Sig	effect size
DAILY LIFE								
SOME LESSONS	-4.5 (3.0)-	-0.08	-4.5 (3.0)-	-0.08	-3.1 (3.3)-	-0.05	-4.5 (3.8)-	-0.07
ABOUT HALF THE LESSONS	-2.1 (3.2)-	-0.04	-2.1 (3.2)-	-0.04	-0.2 (3.5)-	0.00	-2.2 (4.5)-	-0.04
EVERY OR ALMOST EVERY L	0.1 (3.6)-	0.00	0.1 (3.6)-	0.00	1.7 (4.0)-	0.03	-0.1 (4.2)-	0.00
COMPUTER IN CLASS								
SOME LESSONS	2.3 (2.2)-	0.04	2.3 (2.2)-	0.04	0.6 (2.4)-	0.01	2.3 (3.0)-	0.04
ABOUT HALF THE LESSONS	-10.4 (3.8)**	-0.18	-10.4 (3.8)**	-0.18	-11.7 (4.1)**	-0.19	-10.7 (5.1)*	-0.17
EVERY OR ALMOST EVERY L	-20.4 (5.8)**	-0.36	-20.4 (5.8)**	-0.35	-21.0 (6.4)**	-0.34	-20.8 (7.9)**	-0.33
COMPUTER FOR SCH WORK								
A FEW TIMES A YEAR	8.8 (2.4)**	0.16	8.8 (2.4)**	0.15	7.7 (2.7)**	0.12	9.0 (2.9)**	0.14
ONCE OR TWICE A MONTH	4.7 (2.5)-	0.08	4.7 (2.5)-	0.08	5.6 (2.8)*	0.09	4.6 (3.1)-	0.07
AT LEAST ONCE A WEEK	-9.7 (3.6)**	-0.17	-9.7 (3.6)**	-0.17	-7.6 (3.9)-	-0.12	-9.8 (4.2)*	-0.16
EVERY DAY	-15.6 (7.0)*	-0.28	-15.6 (7.0)*	-0.27	-17.3 (7.7)*	-0.28	-15.7 (9.3)-	-0.25

Appendix 6.3. Partitioning of variance and proportion explained

Table 6.3-1: Partitioning of Variance in G4 Mathematics models (MLPV and RUBIN))

G4 MATHS Random Part	MLPV Null Model (VPC)	SE	All Predictors + Controls (VPC)	SE	Proportion of Variance Explained
Level: IDSCHOOL	(0.07)		(0.02)		
	421.3	144.0	90.2	62.0	0.79
Level: IDTEACH	(0.12)		(0.09)		
	709.0	136.9	356.1	74.6	0.50
Level: IDSTUD	(0.69)		(0.70)		
	4025.5	104.6	2706.6	71.5	0.33
Level: IDCASE (PV)	(0.12)		(0.18)		
	703.2	8.5	703.2	8.5	0.00
Total Variance	5859.1		3856.1		0.34

G4 MATHS Random Part	RUBIN Null Model (VPC)	SE	All Predictors + Controls (VPC)	SE	Proportion of Variance Explained
Level: IDSCHOOL	(0.07)		(0.02)		
	428.0	147.7	92.2	65.1	0.78
Level: IDTEACH	(0.12)		(0.09)		
	698.2	141.3	347.0	78.5	0.56
Level: IDSTUD	(0.81)		(0.89)		
	4728.0	118.7	3399.2	85.3	0.27
Total Variance	5854.2		3838.4		0.35

Table 6.3-2: Partitioning of Variance in G4 Science models (MLPV and RUBIN)

G4 SCIENCE Random Part	MLPV Null Model (VPC)	SE	All Predictors + Controls (VPC)	SE	Proportion of Variance Explained
Level: IDSCHOOL	(0.08)		(0.04)		
	466.0	133.4	160.7	64.2	0.66
Level: IDTEACH	(0.10)		(0.07)		
	525.1	113.3	273.1	65.0	0.48
Level: IDSTUD	(0.66)		(0.66)		
	3656.6	97.4	2501.9	68.1	0.32
Level: IDCASE (PV)	(0.16)		(0.23)		
	873.0	10.7	873.0	10.7	0.00
Total Variance	5520.7		3808.7		0.31

G4 SCIENCE Random Part	RUBIN Null Model (VPC)	SE	All Predictors + Controls (VPC)	SE	Proportion of Variance Explained
Level: IDSCHOOL	(0.09)		(0.04)		
	471.9	137.4	161.3	67.5	0.66
Level: IDTEACH	(0.09)		(0.07)		
	508.1	118.0	258.4	69.6	0.49
Level: IDSTUD	(0.82)		(0.89)		
	4569.0	116.2	3358.7	85.3	0.26
Total Variance	5549.0		3778.4		0.32

Table 6.3-3: Partitioning of Variance in G8 Mathematics models (MLPV and RUBIN)

G8 MATHS Random Part	MLPV Null Model (VPC)	SE	All Predictors + Controls (VPC)	SE	Proportion of Variance Explained
Level: IDSCHOOL	(0.07)		(0.05)		
	440.1	294.5	249.2	197.3	0.43
Level: IDTEACH	(0.62)		(0.58)		
	3784.9	404.4	2622.1	282.3	0.31
Level: IDSTUD	(0.23)		(0.26)		
	1421.4	38.7	1165.7	32.2	0.18
Level: IDCASE (PV)	(0.08)		(0.11)		
	495.3	6.0	495.3	6.0	0.00
Total Variance	6141.8		4532.3		0.26

G8 MATHS Random Part	RUBIN Null Model (VPC)	SE	All Predictors + Controls (VPC)	SE	Proportion of Variance Explained
Level: IDSCHOOL	(0.07)		(0.05)		
	434.1	298.3	239.3	197.6	0.45
Level: IDTEACH	(0.62)		(0.58)		
	3798.8	412.1	2588.1	284.7	0.31
Level: IDSTUD	(0.31)		(0.37)		
	1914.0	48.6	1649.9	41.9	0.18
Total Variance	6147.0		4474.2		0.27

Table 6.3-4: Partitioning of Variance in G8 Science models (MLPV and RUBIN)

G8 SCIENCE Random Part	MLPV Null Model (VPC)	SE	All Predictors + Controls (VPC)	SE	Proportion of Variance Explained
Level: IDSCHOOL	(0.31)		(0.17)		
	2013.3	285.9	624.1	98.6	0.69
Level: IDTEACH	(0.08)		(0.06)		
	490.4	79.1	233.9	47.2	0.52
Level: IDSTUD	(0.50)		(0.58)		
	3195.2	91.1	2172.5	62.9	0.32
Level: IDCASE (PV)	(0.11)		(0.19)		
	700.1	8.6	700.1	8.6	0.00
Total Variance	6399.0		3730.5		0.42

G8 SCIENCE Random Part	RUBIN Null Model (VPC)	SE	All Predictors + Controls (VPC)	SE	Proportion of Variance Explained
Level: IDSCHOOL	(0.31)		(0.16)		
	1995.2	287.2	598.1	98.4	0.70
Level: IDTEACH	(0.08)		(0.06)		
	496.1	87.0	230.6	53.9	0.54
Level: IDSTUD	(0.61)		(0.78)		
	3892.3	106.1	2855.1	77.5	0.27
Total Variance	6383.7		3683.8		0.42