ResearchSpace@Auckland

Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognise the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

To request permissions please use the Feedback form on our webpage. http://researchspace.auckland.ac.nz/feedback

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library Thesis Consent Form and Deposit Licence.
Quantitative analysis selection in education

Potential impacts on researchers' conclusions

Kane Lincoln Daniel Meissel

ABSTRACT

This thesis utilises literacy data from a large professional development project in New Zealand schools to investigate the implications of researcher decisions about approaches to quantitative analysis. Specifically, the main research question posed is the extent to which the selection of particular analyses leads to incomplete or even erroneous conclusions when using real-world data. There are various publications advocating for particular methodologies based on this notion, but many of these use simulated data or extreme cases specifically chosen to demonstrate the advantage of the particular method being advocated. This use of simulated and manipulated datasets may explain why many researchers persist with analyses that are argued to be less than ideal. In this thesis, the implications of analysis choice are investigated using increasingly complex analyses, including effect sizes, single-level regression and multi-level models, with the results and conclusions made from each set of analyses collated and compared.

A secondary aim was to evaluate how effectively the professional development project raised student achievement in literacy, and whether these shifts contributed to more equal outcomes for subgroups of students. Much of the previous research where student achievement is linked to professional development has been inconclusive. This inconclusiveness arguably increases the burden of evidence when examining the effect of professional development on student outcomes, so the extensive analyses utilised to investigate the primary thesis aim are especially useful in this regard, and support an in-depth examination of this secondary aim.

The results of the investigations undertaken throughout this thesis showed that the professional development typically resulted in acceleration of progress rates (especially in writing) for all priority subgroups; including students in low socioeconomic catchment areas, and those of minority ethnicities. Choice of analysis was found to be of comparatively minimal importance when considering main effects, but secondary effects were susceptible to sometimes considerable differences in conclusions depending on the method of analysis used. These differences were of sufficient magnitude that policy decisions would likely differ depending on the type of analysis used to infer conclusions.
ACKNOWLEDGEMENTS

As all supervisors, friends, families and partners of those undertaking a doctoral thesis can attest, we doctoral candidates were not the only ones that got strapped into the thesis roller coaster when we optimistically, and perhaps somewhat naively, decided to complete a PhD. They are there with us from that very first moment in which we think of a possible thesis topic, through the many ups and downs and highs and lows of the ensuing years, until that glorious time we all anticipate with ever decreasing expectation of its actual arrival: submission. It is to all of you that I dedicate this thesis, to whom I am immensely grateful for your continued support and positivity, for your depth of understanding and above all, for never letting me fly off the roller coaster.

Special acknowledgement must be made to my supervisors Professors Judy Parr and Helen Timperley for their continued faith in my abilities as well as their insightful comments and guidance.

I am also extremely thankful for the considerable tangible support offered during the course of this thesis. Again, Judy and Helen deserve special acknowledgement for their support in the form of a UniServices scholarship, as well as considerable support to attend conferences and workshops that extended my statistical skills. I am also immensely appreciative of the scholarship provided by Learning Media Limited.

And to those of you who have been especially instrumental in keeping me strapped into the roller coaster, I will be forever grateful. Thanks especially to my wonderful friends, many of whom have already undertaken this journey and have an intimate understanding of its many snares. Thanks also to my family for your love, and for always making me feel like I am doing something worth being proud of. Last, but certainly not least, I wish to express my eternal gratitude to my partner, without whom this would not have been possible. Pau, you are my rock.
# TABLE OF CONTENTS

ABSTRACT ........................................................................................................................................... II

ACKNOWLEDGEMENTS ................................................................................................................... III

LIST OF FIGURES ............................................................................................................................ VII

LIST OF TABLES ............................................................................................................................... VIII

CHAPTER 1. INTRODUCTION .............................................................................................................. 1

  1.1. Research background ............................................................................................................... 2
  1.2. Purpose of the research ............................................................................................................ 4
  1.3. Thesis structure ......................................................................................................................... 7

CHAPTER 2. DATA SOURCES AND CONTEXT .............................................................................. 11

  2.1. The Literacy Professional Development Project (LPDP) ....................................................... 11
    2.1.1. LPDP Design ..................................................................................................................... 14
    2.1.2. Ethical Considerations ....................................................................................................... 16
    2.1.3. Participants ......................................................................................................................... 18
    2.1.4. Tools used in this thesis ..................................................................................................... 20
    2.1.5. Limitations of the LPDP dataset ......................................................................................... 29
    2.1.6. Assumptions made in this thesis ....................................................................................... 32
  2.2. Chapter Summary ..................................................................................................................... 33

CHAPTER 3. EFFECT SIZES ............................................................................................................. 34

  3.1. Standardised mean difference ................................................................................................. 35
    3.1.1. Dealing with dependency .................................................................................................. 38
    3.1.2. Distributional considerations ............................................................................................ 40
  3.2. Alternative effect sizes ............................................................................................................ 42
    3.2.1. Odds ratios ......................................................................................................................... 43
    3.2.2. Cliff’s delta ......................................................................................................................... 44
  3.3. Analysis .................................................................................................................................... 47
    3.3.1. Analysis: Odds ratios ........................................................................................................ 48
    3.3.2. Analysis: Standardised Mean Difference versus Cliff’s delta ........................................... 52
  3.4. Chapter Summary .................................................................................................................... 68
6.2.2. LPDP success at improving writing achievement and reducing group differences........184

6.3. Concluding statement.......................................................................................................185

REFERENCES..........................................................................................................................187
LIST OF FIGURES

Figure 1. Typical pattern of curriculum level achievement by school year level (The New Zealand Curriculum Online, 2007). .................................................................22

Figure 2. Proportion of distributional overlap for Cohen’s d = 0.5 ........................................47

Figure 3. Histogram of T1 Score against expectation for Cohort 2 Writing ......................53

Figure 4. Normal Q-Q Plot of T1 Score against expectation for Cohort 2 Writing ............54

Figure 5. Basic regression equation showing the average relationship between baseline and EOY1 asTTle reading scores for a 10% sample of Year 7 students in the 2008-2009 cohort. ........................................................................................................................................79

Figure 6. Various regression equations describing the relationship between baseline and EOY1 asTTle reading scores for a 10% sample of Year 7 students in the 2008-2009 cohort of LPDP ..............................................................................................................................86

Figure 7. Level 1 regression line for an individual school ..................................................126

Figure 8. Level 2 regression lines for individual schools with varying intercepts but fixed slopes ..............................................................................................................127

Figure 9. Level 2 regression lines for individual schools with varying intercepts and slopes. ..............................................................................................................128

Figure 10. Differences in average value added by facilitators for reading .......................169

Figure 11. Differences in average value added by facilitators for writing .......................171

Figure 12. Difference between the estimated progress for each subgroup from the overall cohort gains (the zero line), showing the range of estimates across analysis type ....177

Figure 13. Estimated gains made by subgroup in the 2006-2007 reading cohort by procedure. ..............................................................................................................179

Figure 14. Estimated gains made by subgroup in the 2008-2009 reading cohort by procedure. ..............................................................................................................180

Figure 15. Estimated gains made by subgroup in the 2006-2007 writing cohort by procedure. ..............................................................................................................181

Figure 16. Estimated gains made by subgroup in the 2008-2009 writing cohort by procedure. ..............................................................................................................182
LIST OF TABLES

Table 1. Comparison of Ethnic Composition of the LPDP Sample Against asTTle Normative Sample. ........................................................................................................................................................................20

Table 2. The odds of students with baseline scores in the lowest 20% progressing above asTTle expectations for their year group by subgroup. ..................................................................................................................51

Table 3. Descriptive statistics for the first year of the project: Reading 2006-2007. .............57

Table 4. Descriptive statistics for the second year of the project: Reading 2006-2007 ........58

Table 5. Comparison of NPSMD and SMD estimates of progress effect sizes in each year of the project, and for the project overall: Reading 2006-2007 ........................................................................................................59

Table 6. Descriptive statistics for the first year of the project: Reading 2008-2009. ..........60

Table 7. Descriptive statistics for the second year of the project: Reading 2008-2009 ..........61

Table 8. Comparison of NPSMD and SMD estimates of progress effect sizes in each year of the project, and for the project overall: Reading 2008-2009 .........................................................................................................62

Table 9. Descriptive statistics for the first year of the project: Writing 2006-2007............63

Table 10. Descriptive statistics for the second year of the project: Writing 2006-2007 ......64

Table 11. Comparison of NPSMD and SMD estimates of progress effect sizes in each year of the project, and for the project overall: Writing 2006-2007 ....................................................................................................65

Table 12. Descriptive statistics for the first year of the project: Writing 2008-2009..........66

Table 13. Descriptive statistics for the second year of the project: Writing 2008-2009 .......67

Table 14. Comparison of NPSMD and SMD estimates of progress effect sizes in each year of the project, and for the project overall: Writing 2008-2009 .............................................................................................68

Table 15. LPDP added value within reading by subgroup after removing estimated RTM effect. ........................................................................................................................................................................95

Table 16. LPDP added value within writing by subgroup after removing estimated RTM effect. ........................................................................................................................................................................96

Table 17. Final stepwise regression model showing the effect of each factor on EOY2 score against expectation (n=4,163) among students in the 2006-2007 reading cohort. ..............100

Table 18. Final stepwise regression model showing the effect of each factor on EOY2 score against expectation (n=1,510) among students in the 2008-2009 reading cohort. ..............102
Table 19. Final stepwise regression model showing the effect of each factor on EOY2 score against expectation (n=3,700) among students in the 2006-2007 writing cohort. ............ 104

Table 20. Final stepwise regression model showing the effect of each factor on EOY2 score against expectation (n=3,043) among students in the 2008-2009 writing cohort. ............ 106

Table 21. Consistency of the shifts among the reading schools working with facilitators with lower added value. ............................................................................................................ 109

Table 22. Consistency of the shifts among the reading schools working with facilitators with higher added value. ............................................................................................................. 110

Table 23. Consistency of the shifts among the writing schools working with facilitators with lower added value. ............................................................................................................ 111

Table 24. Consistency of the shifts among the writing schools working with facilitators with higher added value. ............................................................................................................. 112

Table 25. Number of data points at each level of the hierarchical linear models. ............ 142

Table 26. Unconditional models comparing variance proportion at each level by estimation procedure. ............................................................................................................. 147

Table 27. Final HLM showing parameter estimates by estimation method for 2006-2007 reading cohort. ............................................................................................................. 149

Table 28. Remaining variance and variance explained by estimation method. ............ 152

Table 29. Unconditional models comparing variance proportion at each level by estimation procedure. ............................................................................................................. 154

Table 30. Final HLM showing parameter estimates by estimation method for 2008-2009 reading cohort. ............................................................................................................. 155

Table 31. Remaining variance and variance explained by estimation method. ............ 157

Table 32. Unconditional models comparing variance proportion at each level by estimation procedure. ............................................................................................................. 158

Table 33. Final HLM showing parameter estimates by estimation method for 2006-2007 writing cohort. ............................................................................................................. 160

Table 34. Remaining variance and variance explained by estimation method. ............ 162

Table 35. Unconditional models comparing variance proportion at each level by estimation procedure. ............................................................................................................. 163
Table 36. Final HLM showing parameter estimates by estimation method for 2008-2009 writing cohort.

Table 37. Remaining variance and variance explained by estimation method.
CHAPTER 1. INTRODUCTION

This thesis investigates the extent to which educational researchers’ choices about how to analyse quantitative data may conceal vital aspects of the story told by the data, resulting in researchers making incomplete or even misleading and erroneous conclusions. This is particularly relevant in the field of educational research, a research field in which there is substantial public (and government) engagement, and where research contexts (i.e., schools) are highly complex (Berliner, 2002; Paterson & Goldstein, 1991), requiring considerable expertise and care to ensure that evidence derived is couched within appropriate quantification and analytic methods. There has been extensive commentary in the literature lamenting the failure of many educational researchers to address the appropriateness of the analyses used adequately, or even to check whether their data meet the basic assumptions of the statistical tests being reported (i.e., Ahn, Ames, & Myers, 2012; Costello & Osborne, 2005; Gorard, Rushforth, & Taylor 2004; Henson, Hull, & Williams, 2010; Hess & Kromrey, 2004; Osborne, Waters & Waters, 2002). Although these (and other) authors have been raising concerns about possible implications for the conclusions made by researchers for well over a decade, this situation persists.

There are likely various reasons why researchers might fail to follow best practice with respect to quantitative analysis (and journal reviewers continue to accept such work), not the least of which is a general lack of quantitative expertise (Gorard et al., 2004; Costello & Osborne, 2005; Henson et al., 2010), and many of these reasons are beyond the scope of this thesis. However, much of the previous work espousing the need for improved quantitative methodology has discussed possible implications from a predominantly theoretical perspective; either by reviewing extensive published literature and identifying possible issues with the analyses presented (i.e., Henson, Capraro & Capraro, 2004; Keselman et al., 1998; Micceri 1989; Zientek, Capraro & Capraro, 2008), using simulated data with properties specifically selected to demonstrate the advantage of one method over another (e.g., von Oertzen, Ghisletta, & Lindenberger, 2010), or by positing the process required for best practice analysis (the focus of the journal Practical Assessment, Research and Evaluation, or PARE). Conversely, there has
been a dearth of literature investigating the potential ‘real-world’ implications of the quantitative methodology gap. Little research has been presented that demonstrates the possible effects of neglecting to follow best practice within applied contexts, using real data that applied researchers can relate to and understand. While there are some recent contributions where both simulated and real data are used (there has been a notable increase in recent issues of the Journal of Educational and Behavioral Statistics; e.g., Feldman & Rabe-Hesketh, 2012; Lee & Cai, 2012), the very high level of statistical knowledge required to comprehend these articles ensures that these articles remain firmly targeted at researchers who already have considerable quantitative expertise. This thesis aims to provide a more generally accessible contribution to this debate, utilising unaltered literacy data from a large professional development project undertaken in New Zealand schools to investigate how researcher decisions regarding their data analysis may have implications for results and subsequent conclusions (and, potentially policy decisions).

1.1. Research background

In recent years, there has been an increased focus on evidence-based education, in part arising from considerable debate about the perceived lack of quality and scientific rigour of educational research, as well as a response to increased political and ideological demands to determine what actually works for students (e.g., Bennett, Lubben, Hogarth, & Campbell, 2005; Davies, 1999; Feuer, Towne, & Shavelson, 2002; Gorard, 2001; Gorard, et al., 2004; Hammersley, 2005; Henson et al., 2010; Oakley, Gough, Oliver, & James, 2005; Yates, 2004). This focus has led to renewed attention on the methods used to conduct and analyse quantitative research, frequently drawing on methods used in medicine and other sciences, due to a perception that such methodologies provide increased rigour and greater generalizability (which is of particular political significance; e.g., Feuer et al., 2002; Levin, & O'Donnell, 1999). The effectiveness of this shift has been questioned however, with critics of this renewed focus expressing concern about attempts to “empiricise” a field that many argue bears little relationship to the laboratory, thereby limiting the applicability of methodologies derived for this latter purpose.
(e.g., Hammersley, 2005a; Hammersley 2005b; Hodkinson, 2004). Such critics further assert that the evidence-based movement negates and neglects "evidence" derived through experience, resulting in a narrow, reductionist perspective of evidence. Despite these arguments, there is little real justification for treating qualitative and quantitative methodologies as separate and distinct (Feuer et al., 2002; Gorard, 2010; Onwuegbuzie & Leech, 2005; Tashakkori & Teddlie, 1998) and, instead, judicious combination of available methods should be seen as an effective way of conducting relevant and high quality research (Feuer et al., 2002; Gorard & Cook, 2007; Cook & Gorard, 2007). Indeed, many researchers argue for research method pluralism (e.g., Bryman, 2005; Feuer et al., 2002; Johnson & Onwuegbuzie, 2004; Onwuegbuzie & Leech, 2005; Rey, 2006), arguing that particular methodologies should not be at the expense of other approaches and should, rather, be seen as complementary.

This is not a unanimous view however; even quantitative advocates are sometimes unconvinced about how effective the evidence-based movement has been, raising concerns about the piece-meal and frequently ad-hoc approach to quantitative analysis evident in education (e.g., Gorard, 2001; Gorard et al., 2004). These researchers argue that such concerns arise as a result of the lack of skills of many researchers with respect to quantitative methodologies, which results in the acceptance without critique of such analyses, selective sampling only of research that offers support for researchers’ existing ideologies, or active disregard of quantitative results. Researchers (understandably) tend to focus on the particular research question of interest, often leaving decisions about which methods to use for analysis and interpretation out of the design and planning stage. However, this lack of attention to analysis methodology is argued to result in such decisions frequently being based on factors that should be unrelated, such as the particular skillset of the individual researcher (Gorard et al., 2004), in combination with a bias toward simply selecting the default option (default option bias) of the particular software package being used for analysis (Costello & Osborne, 2005), rather than what is most appropriate for the research question. While it makes sense that researchers will show bias toward methods they have used in the past, blind faith in the robustness of such approaches to analysis has little place in an academic community.
The default option bias is equally concerning. Costello and Osborne (2005) suggest that the popularity of certain procedures is largely due to a tendency for researchers to resort to the default option within statistical software packages when unsure about the most appropriate technique to use. These authors contend that an example of this bias is the popularity of factor analyses using principal component analysis (PCA) as the extraction method (Costello & Osborne, 2005). Further, it is argued that the underlying assumptions of PCA are typically inappropriate for the educational context, and should therefore generally be avoided. For example, PCA techniques assume there is no measurement error, which is generally not the case in educational research. Consequently, Costello and Osborne suggest that the main reason for the popularity of PCA is that it is the default extraction method in statistical software suites such as SPSS and SAS. This tendency to resort to the default option raises questions about the comparability of results if analyses are so frequently determined by the default option, rather than what is most appropriate for the research question or data properties.

1.2. Purpose of the research

The primary aim of this thesis is to investigate the extent to which analysis choice has implications for conclusions made about a given dataset. It is not an exercise in applying erroneous methodologies; rather, it involves consideration of a range of quantitative analyses that are commonly applied in education, coupled with commentary and investigation into how these choices affect conclusions made about the results. This aim is especially important since it focuses on analytic methodology in an applied sense and helps to highlight the rather different conclusions that may be made about particular interventions simply due to researcher decisions or preferences for a particular method of analysis, rather than as a result of real and meaningful differences in the data. It is hoped that this thesis will contribute to understanding around statistical analysis methodologies used within the educational context, and provide other researchers with a deeper awareness of the possible consequences of the choices they make.
with respect to analysis. In turn, it is hoped that this will contribute to the way in which researchers approach future research questions.

A secondary aim is to provide an evaluation of the effectiveness of the nationally funded professional development project which aimed to improve student achievement in literacy (reading and writing), and from which the data used in this thesis were drawn. The effectiveness of this project will be assessed through consideration of the results of the analyses conducted in addressing the primary aim. The data collection process was completed prior to the commencement of this thesis so is not included as a substantive component of the work presented. However, in order to provide adequate contextualisation and reduce repetition, the nature of the project and its methodology, data and data collection processes, are described fully in Chapter 2. The project previously has been described as very successful in its research reports, but has largely been evaluated using fairly simplistic analyses (such as pairwise effect sizes); partly in an attempt to ensure the results were accessible to practitioners. Of particular interest are the factors that support or attenuate literacy progress, both at the individual level (demographic variables such as gender, ethnicity and initial achievement) and at the school level (collective socio-economic or ethnic composition), within the context of a professional development intervention. Developing a better understanding of factors that explain the nature of progress and distribution of success among students within a national professional development project is essential for determining what works, especially with respect to areas in which the intervention has been more or less successful in supporting greater progress among students most in need. In doing so, it provides information essential to the iterative feedback process outlined by Timperley and Parr (2009), and incorporated in the project. In turn, this may assist subsequent interventions to be increasingly meaningful for student achievement and progress and will also likely be useful to researchers attempting to implement similar professional development in their own context.

In order to address the secondary aim of evaluating the success of the professional development project described in this thesis, analysis procedures must be comparable across different projects. Comparability is one of the stated aims behind the use of measures of effect
size – and indeed this is the premise behind the use of effect sizes in meta-analysis (Ahn et al., 2012; Hattie, 2009). However, a recent review of more than 50 meta-analyses in education published since the year 2000 concluded that considerable deficits remain with respect to the quality of data analysis employed by educational researchers – both those conducting the original studies, as well as those conducting the meta-analyses (Ahn et al., 2012). This review showed that researchers frequently report very different information about the analyses conducted, with many reporting insufficient information for accurate comparison. While many studies reported effect sizes, many failed to include basic details about distributional properties, such as means or standard deviations, or whether normality assumptions were met. In addition, many neglected to assess whether regression to the mean was a possible confound, or failed to even identify which effect size was presented, making it extremely difficult to assess relative magnitude of the effect sizes of different studies.

Almost two-thirds of the meta-analyses included in the review reported having encountered multiple studies where dependency in the data was ignored (a fundamental assumption of commonly used effect sizes and meta-analytic aggregates is that the data are independent), while the remainder of authors failed to report whether they had accounted for the data dependency, despite having included multiple studies where dependency was clearly an issue. Failure to account for dependency by ignoring clustering or hierarchical levels in the data (i.e., the notion that individuals within a group are more likely to be correlated with each other than with individuals in another group), is argued to seriously jeopardise any conclusions from such studies due to the risk of atomistic ¹ and ecological ² fallacy (Goldstein, 1995; Hox, 2010). Ahn and colleagues (2012) conclude that these inconsistencies and oversights raise serious concerns about the validity of the conclusions being made from these meta-analyses. The primary aim of the thesis contributes to this debate by investigating the extent to which conclusions differ using different analyses on the same dataset – if there are substantive differences within the same

¹ Making inferences about groups based on patterns observed among the individuals belonging to particular groups
² Making inferences about individuals based on patterns observed in groups to which the individual belongs
dataset, this raises serious concerns when attempting to compare success across different datasets.

Reasons for the oversights indicated above are likely to be complex and, realistically, no two studies are precisely comparable due to differences in both aim and method – and indeed appropriate methods differ depending on the aims of the research. However, if successful interventions are to be identified, replicated, and adopted, it is imperative that researchers use methods and analyses that support comparability and replication. This thesis aims to contribute to awareness about the importance of using best practice recommendations for analysis purposes, by investigating the potentially differing conclusions reached by making different decisions about which analysis to use. The series of analyses presented will also allow examination of the various determinants of student progress in the context of a professional development project (LPDP).

The overarching research question for this thesis is:

- Does choice of analysis have a substantive effect on conclusions made about the results when analysing a large educational dataset?

Additional research questions for this thesis are:

- Did the professional development provided in LPDP result in gains in student achievement?
- Were there differential effects among subgroups of students?

1.3. Thesis structure

This chapter provided the contextual framework for the thesis. As indicated above, Chapter 2 will provide contextual information about the professional development project from which the data analysed throughout this thesis were drawn. In addition, a brief discussion of the research used to inform the professional development project is presented. Since the primary focus of the thesis relates to the consequences surrounding choice of analysis rather than professional development specifically, this is not exhaustive. Rather, it is intended that this section will help to
provide the reader with sufficient detail to understand the rationale and context within which LPDP sat.

A general picture of the overall effectiveness of LPDP is presented in Chapter 3 by using single-level pair-wise group difference analyses, or effect sizes. It has become virtually a requirement to report effect sizes following recommendations by the American Psychological Association (APA, 2001) and the American Educational Research Association (AERA, 2006), with the APA stating that “it is almost always necessary to include some index of effect size or strength of relationship” (APA, 2001, p. 24) and AERA asserting that “interpretation of statistical analyses is enhanced by reporting magnitude of relations (i.e., effect sizes) and their uncertainty separately” (AERA 2006, p. 37). Since these recommendations, the proportion of researchers reporting an effect size has increased markedly, yet concerns remain about the appropriateness of the effect sizes typically presented (Peng, Chen, Chiang & Chiang, 2013). There are numerous effect size measures available based on various properties of the data, with group differences measured in standard deviation units (e.g., Cohen’s $d$, Glass’s delta, Hedge’s $g$) by far the most commonly used measure of effect within education. The results of the 800 meta-analyses reported in Visible Learning (Hattie, 2009) were aggregated using Cohen’s $d$. However, the popularity of such measures is arguably a result of their ease of calculation rather than any particular suitability for the data being analysed, so Chapter 3 includes a discussion of the various effect metrics that are commonly used, along with a comparison of the relative performance of commonly used metrics against an unbiased but lesser known non-parametric alternative.

In Chapter 4, single-level regression models are developed. This mirrors the progression shown in the analyses of Success for All (SFA), another large-scale literacy intervention, which has been widely adopted internationally. Initially, much of the research focused on effect sizes as an indication of efficacy (e.g., Slavin, Madden, Kanweit, Livermon & Dolan, 1989; 1990), in much the same way as reported in the LPDP milestones. However, more recent evaluations have moved toward more complex analysis, while continuing to use effect sizes as supplementary information. For example, Tymms and colleagues have conducted extensive research where baseline achievement is taken into account, providing a “value-added” metric, typically derived
using a form of single-level linear regression (e.g., Tymms, 1999; Tymms, Merrell, & Jones, 2004). This methodology was used to evaluate the SFA program (Tymms & Merrell, 2001), and is argued to provide a fairer measure of whether a teacher, school or program has added something differentially greater than others in the same category, since initial achievement is taken into account within the model. It also allows covariates to be considered within one model so that the relative contribution to the total effects can be considered separately, while controlling for the contribution of the other covariates included in the model.

However, some argue that even value-added research is unreliable due to the immense range of possible explanations for differences (e.g., Gray, Jesson, & Jones, 2002). In addition, single-level models carry the risk of atomistic and ecological fallacy where assumptions are made about individuals or groups that are not possible from the analysis, which can lead to conclusions that are the opposite to what the data actually indicate (this is discussed in Chapter 5). More recently, the techniques used to analyse the effectiveness of the SFA program have shifted toward multi-level modelling which incorporates the system complexities evident in educational contexts by treating data as hierarchical, with, for example, students nested within classrooms within schools (e.g., Borman et al., 2005a; Borman et al., 2005b). These modelling techniques have been available for at least 30 years, but uptake has been relatively limited due to computational (initially) and researcher knowledge limitations, along with the practical requirement for a considerable amount of data - and indeed, this process took 15 years in the Success for All program. Borman and colleagues (2005a; 2005b) indicate that the shift to multi-level modelling arose out of a desire to incorporate as much contextual detail as possible into their SFA models, and to take advantage of the large amount of data available, to provide a more accurate project evaluation. Multilevel models are presented for the LPDP data in Chapter 5.

This thesis follows an iterative methodology, with each chapter presenting results using increasingly complex techniques, allowing for both local and holistic consideration of the results and conclusions derived from the various analyses. In each chapter, the same set of factors is always investigated; at a minimum, these factors are gender, ethnicity and socio-economic status of the school. It is intended that this will maximise comparability across analyses and facilitate
conclusions about whether the analysis choice would have an impact on researchers’ conclusions with these data (i.e., the primary aim of the thesis). Other factors are also included for certain analyses when the specific methodology allows, and when inclusion of these factors provides substantive additional information for either research aim. The final chapter (Chapter 6) summarises and aggregates the conclusions made about the professional development project based on the results of each of the statistical methods, and identifies and discusses where the results of each set of analyses are the same or similar, and where they differ.

This thesis follows a similar progression with respect to the analyses used, as those used to evaluate the SFA program, but goes further by explicitly focusing on and investigating the differences in conclusions drawn from the various results of different analyses. Borman and colleagues (2005a; 2005b) comment that the main school-level results were comparable to those obtained in earlier research based on simpler analyses, but do not make any comparison at the student level, since their primary focus was assessing the project’s success. The effectiveness of the professional development project assessed in the current context (the secondary aim of the thesis) is addressed by making conclusions based on an holistic consideration of the aggregation of all results, with a weighting toward those deemed more appropriate for the data. This aggregation of results also provides a base for the discussion of the primary research question – to what extent does it matter which analysis researchers choose? To the best of my knowledge, there has not been an extensive consideration of the implications of analysis choice using such a large, (non-simulated) nationally representative dataset as that presented in this thesis. Given the size and quality of the dataset used in this thesis, any differences identified are likely to be meaningful, with important implications for other educational researchers whose datasets are typically much smaller and therefore, at even greater risk of misinterpretation due to inappropriate analysis choices.
CHAPTER 2. DATA SOURCES AND CONTEXT

The introductory chapter detailed the research questions for this thesis, as well as providing some detail on the motivation for these investigations. Chapter 1 also noted that the investigations detailed in this thesis draw on longitudinal data from a large, nationally funded, professional development project. The current chapter initially provides an introduction to this project to assist with an understanding of the context in which this thesis is framed, followed by an overview of the project design. This overview is followed by a discussion of the ethical issues related to the use of the data from this project in section 2.1.2. Details about the participants for whom data were provided are included in section 2.1.3. The following section provides details about the tools used in the professional development project, with a particular focus on the assessment tools used in the analyses undertaken in this thesis. Section 2.1.5 discusses the limitations of the dataset, while the following section indicates the major assumptions made about these data when undertaking the analyses detailed in this thesis.

2.1. The Literacy Professional Development Project (LPDP)

The Literacy Professional Development Project (LPDP) was introduced as part of the New Zealand Ministry of Education’s strategy to improve student progress and reduce disparities among subgroups (Ministry of Education, 2006), particularly with respect to gender, socio-economic and ethnic groupings (OECD, 2001; 2005). Although the average scores for New Zealand students compare consistently well in international studies such as PIRLS (Mullis, Martin, Kennedy & Foy, 2007) and PISA (OECD, 2001; 2005), the disparities among subgroups, especially those delineated by ethnicity, compare less favourably. Internationally, much of the response to such disparities focused on school reforms, centring primarily on (re-)educating and developing teachers, since teachers are considered to have the most impact on student achievement at the system-level (Alton-Lee, 2003; Nye, Konstantopoulos, & Hedges, 2004). The project was derived from this premise and, in turn, was based on the hypothesis that sufficiently
targeted improvements in pedagogical practice would transfer to the student level, resulting in measurable gains in student achievement (Bareta, English & O’Connell, 2006). There was a particular focus on four outcomes: evidence of improved student achievement; evidence of improved teacher content knowledge; evidence of improved transfer of understanding of literacy pedagogy to practice, and evidence of professional learning communities. The first outcome was a way of measuring whether the strategy was successful, while the others were intended to facilitate and promote the conditions necessary to achieve the ultimate goal of disparity reduction.

An additional complexity moderating the extent to which a professional development program is successful, is the considerable diversity of both teachers and students in terms of ethnicity, ability and schooling context and, for teachers, age and experience (Timperley, Fung, Wilson & Barrar, 2007). These contextual features complicate the implementation of professional learning, as different sets of teachers and students will have differing needs. The research team reports taking these differences into account by providing a collaborative analysis of teachers’ learning needs with the teachers and leaders of each school, prior to the delivery of any professional development (Parr, Timperley, Reddish, Jesson & Adams, 2007), focusing on both how teachers learn, as well as what they need to learn. The needs analysis allowed the professional learning to be based on clear conceptual frameworks that take account of conceptions of students and how they acquire knowledge, to pinpoint the knowledge that teachers value, as well as teacher beliefs about the pedagogical practices that are most effective. This is argued to be essential, since teachers do not implement changes to their practice unless it sits within the context of their existing beliefs and the realities of their classrooms (Donovan, Bransford, & Pellegrino, 1999; Kennedy, 2004; Robinson, 1993; Robinson & Lai, 2006).

A further factor that affects teacher effectiveness is the socio-organisational context of the school, so LPDP focused on the specific learning needs of teachers within the context of their school (Parr, Timperley, Reddish, Jesson & Adams, 2007). In addition, the professional development supplied to school leaders focused on the role of literacy leaders and principals due to increased awareness that such positions require considerable pedagogical content knowledge.
in addition to simple organisational ability (Robinson & Lai, 2006). Without this pedagogical knowledge, leaders cannot facilitate teacher professional learning.

The formulation and placement of LPDP within the literature provides an interesting context within which to examine the features of professional development that potentially generate substantive gains for students. LPDP followed a collaborative pathway, with expert facilitators working with school leaders and teachers to identify and address needs, coupled with considerable focus on developing teacher and leader abilities to self-regulate their learning. This ability to self-regulate was considered essential as it would allow leaders and teachers to become adaptive, and control their own learning even after the completion of the project. Project milestones for this project indicated that students whose schools were involved made considerably more progress than a broadly comparable normative sample during each of the three, two-year cohorts. Even more importantly, the milestones note that gains were greatest among the lowest achieving students, with gains in writing five to six times the expected rate of progress and gains in reading more than three times the expected rate (Timperley, Parr & Meissel, 2010). In a pseudo-randomly selected sample of schools, most sustained the rate of gain for new student cohorts for the entire follow-up monitoring period of three years (O’Connell, 2010), suggesting that the changes in practice were sustained and continued to be effective.

There is considerable variability evident in the literature about the extent to which professional development has an effect on student-level outcomes (Timperley et al., 2007), so developing further understanding of the features that make an intervention successful is essential. LPDP reported large gains in achievement in its project milestones, and credited these gains to coherence within and between the multiple levels of the schooling and administration systems, in combination with the development of educational partnerships alongside a focus on evidence-informed inquiry into effectiveness at each level of the system (Parr et al., 2007; Parr & Timperley, 2010; Timperley et al., 2010). Although the primary focus of this thesis relates to the implications of using different quantitative methodologies, consideration of the complete set of analyses also provides strong evidence of the extent to which LPDP was successful. It is hoped that this will
assist other researchers determine whether there are lessons that may be gleaned from the LPDP experience.

2.1.1. LPDP Design

The design of LPDP was collaborative, in that the project team included staff from both Learning Media Limited (LML) and the University of Auckland, as well as a representative from the New Zealand Ministry of Education. LML was responsible for the management of the project and in-school delivery of the professional development via expert facilitators who each worked closely with a small number of schools. Staff at the University (the research team) worked collaboratively with LML, and provided specific and ongoing feedback based on evidence being generated from the project.

The project ran from 2004-2009; with more than 300 schools electing to take part during this period. Participation was voluntary, and each school chose which particular aspect of literacy they wanted to focus on; reading, or writing. Typically schools that chose to focus on writing showed baseline evidence of greater need for additional support than those that chose to focus on reading. There were three main cohort intakes, of two years each, beginning in 2004, 2006 and 2008. The data were collected using a quasi-experimental methodology, with professional development being introduced to schools, and the subsequent gains being compared against a broadly comparable normative sample. Schools were able to choose whether to focus primarily on either reading or writing, and achievement data were collected for the specific focus chosen.

Various baseline data were collected as part of the needs analysis in order to gain a clear snapshot of the strengths and needs of each school in the project. These variables included: baseline student achievement data; relevant teacher pedagogical content knowledge; knowledge of how to use evidence and interpret data; teacher beliefs and confidence in setting learning objectives; measures of school organisation and climate with respect to innovations; and observations of current practice, among others. Much of these data formed an essential part of the delivery of LPDP, but the fundamental goal was to improve student outcomes. Only the
student achievement data were collected at three time points since these data were considered essential to the evaluation of the project; the beginning of the first year (baseline), again at the end of the first year, and then the third and final assessment at the end of the second project year. In addition, only the student achievement data were collected using comparable measures with distributional properties that were appropriate for the range of analyses conducted. As a result, these other measures are not described, but complete details are available in LPDP milestone reports (e.g., Parr et al., 2007).

There were approximately 25 expert facilitators working nationally at any one time (the number varied depending on the size of the cohort, the number of students in each school, and the geographical distribution of the participating schools), with each facilitator working with a small number of schools in their assigned area. These facilitators were also supported by a development team comprising eight educational consultants and two researchers, with a considerable focus on facilitator learning. This related to the addition of a contracted outcome of establishing whether there was evidence of effective facilitation (Bareta et al., 2006). This stemmed from previous research suggesting a requirement for facilitators to make considerable knowledge gains, and to adapt their practices to schools in order to produce effective transfer and sustained growth in student-level outcomes (Timperley, Parr & Higginson, 2003). Since the facilitators tailored each school’s professional development around the needs analysis produced in collaboration with teachers and leaders at baseline, the development curriculum was adaptive rather than prescriptive, based on the principle of concept dissemination discussed by Datnow, Hubbard and Mehan (2002; Datnow et al., 2003). The role of the facilitators was one of scaffolding, based on co- and self-regulated inquiry via systematic, data-informed reflection. It was intended that this support would provide teachers and leaders with the knowledge required ultimately to provide support and professional development for each other.
2.1.2. Ethical Considerations

The data analysed in this thesis were collected during the 2006-2007 and 2008-2009 cohorts of the LPDP research programme undertaken by The University of Auckland and Learning Media Limited. Ethics approval for LPDP was granted by The University of Auckland Human Participants Ethics Committee (UAHPEC) on 13 May 2004, with a further extension granted for 2007–10 (reference no. 2004/059). The application stipulated that pre and post-intervention student achievement data would be collected, analysed and linked with the various other sources of information derived by LPDP in order to develop evidence-based practice. Boards of Trustees at each school entered into a memorandum of understanding with Learning Media Limited, which gave permission for the data to be used.

The current project falls under the umbrella of the original and extended ethics applications; no additional data were collected during the thesis, and the analyses contribute to the fulfilment of the aims outlined in the original application in the sense that they provide evidence of the extent to which the project effected change in student achievement. The primary aim of the thesis, which is to investigate the effect of making different decisions about analysis, is essential to addressing the original project aims adequately since it assesses the extent to which the results hold across different analyses. Data have been treated according to the conditions outlined within the application; no individual or school is identified and data will be discarded six years after the expiration of the ethical approval (in 2016) as stipulated by the ethics committee. Care has been taken when reporting identifying features (e.g., ethnicity, gender etc.) to ensure that this is done in a way that does not allow identification of individuals or schools.

The researcher was a member of the LPDP research team for three years prior to enrolment as a PhD candidate at The University of Auckland, working as the data analyst. In this role, I signed a confidentiality and disclosure agreement, and as a PhD student receiving scholarship funding from UniServices Limited, signed an additional confidentiality agreement requiring all dissemination of information to be first approved by the thesis supervisors. In addition, I was added as a named researcher in an addendum to the ethics application, on the 14th of October, 2011, approved by the acting chair of the ethics committee on the 25th of October, 2011.
Although the ethics approval endorses the data collection procedures, this thesis is primarily a methodological and analytical exercise, and must also be conducted according to the ethical principles identified by the UAHPEC. Principles (from the UAHPEC Applicants’ Manual, 2013) considered relevant to this dimension of the research, and not already discussed, are identified and addressed below.

Conflicts of interest

The researcher held two scholarships during the period of this thesis, one from Learning Media Limited and another from UniServices. The Learning Media scholarship was provided for research investigating the use of teaching materials in literacy (such as those used in LPDP), but did not dictate methodology or results. The UniServices scholarship was similarly provided without any conditions with respect to methodology or results and at no time did the funders of either scholarship put pressure on the researcher to present results in a certain way. However, it could be argued that there is a possible conflict of interest since the researcher could be perceived to have a vested interest in the results due to receiving funding to undertake the research. This conflict is mitigated since the scholarship funding was provided for a period after LPDP had already been completed.

Publication of results

One of the primary goals within the course of this PhD research was to develop publishable articles. In doing so, there is concern around the use to which findings may be put – particularly within an educational context where findings have the potential to influence policy and/or perceptions about sub-groups of students. To address this, care has been taken to ensure that the findings are as comprehensible as possible to all potential readers, and possible impact on specific groups has been discussed with the thesis supervisors, and advice sought from Māori and Pasifika advisors within the Faculty of Education.
Issues related to the Treaty of Waitangi

As stated in the UAHPEC guidelines, research including Māori participants must build partnership and equity between the researcher and Māori. Within the context of this thesis, this means that the outcomes of the research need to directly or indirectly benefit Māori. In order to ensure that this is the case, findings have been related and discussed in terms of how they may be used beneficially (e.g., to improve progress, or interventions aimed at raising achievement and/or progress). The notion of ensuring that the outcomes of the research are beneficial has also been a general principle of this thesis; understanding the possible limitations and complexities that may be introduced by selection of different analyses is an essential component of developing equitable outcomes. This applies equally to discussion around findings related to any group – especially those applying to minority or underachieving groups.

2.1.3. Participants

To ensure maximum comparability across the various analyses conducted in this thesis, only achievement data from the Assessment Tools for Teaching and Learning tool (asTTle; this tool is explained in the following section) from the 2006 and 2008 cohorts are included. This decision was made because the asTTle data were gathered in the most reliable and consistent way, and most students were tracked using this tool. In addition, of all the achievement measures used, only asTTle had the distributional properties required for the wide range of analyses being conducted in this thesis. The number of students included in the analyses presented throughout the thesis differs somewhat depending on the constraints of the particular analysis – especially with respect to treatment of missing data. Therefore, specific numbers included in the analysis are noted in each chapter, but overall totals, aggregated by focus, are also provided here.

Among schools with a reading focus, longitudinal data were available for 8,359 students (i.e., they were present for at least two of the three time points) attending 115 schools, while for writing, longitudinal data were available for 11,749 students attending 86 schools. For reading, this includes all schools in the 2006-2007 cohort of LPDP, and approximately two-thirds of the 2008-2009 cohort, while for writing, all schools that provided data were included in the analyses.
presented in this thesis. The reason for this difference was that in the latter cohort, some schools with a reading focus opted to use the Progressive Achievement Test (PAT; explained briefly in the following section) to measure achievement and progress, meaning that these results were not comparable with those derived using asTTle. Since the vast majority of data were collected using asTTle, and due to the analytic flexibility afforded by the distributional properties of this tool, asTTle has been retained as the measure of interest.

The socioeconomic profile of schools that chose PAT over asTTle was higher than for other schools. In New Zealand, schools are assigned a decile rating, which is an aggregate measure dividing schools into ten groups based on the degree to which a school draws its students from particular socio-economic areas. A decile rating of one indicates that the school is among the 10% of schools with the highest proportion of students from low socio-economic backgrounds. The decile is calculated by the NZ Ministry of Education in five-year cycles (after each census), and is based on household income and crowding, parental occupation and qualifications, and degree of income support. The decile assigned to a school has funding implications, with additional funding provided the lower the socio-economic background of its students. Schools choosing PAT as the assessment tool were far more likely to have a high decile rating (8-10), but despite this, the proportion of students in each decile band was broadly similar to the national picture among students attending schools that used asTTle.

Certain demographic details were similar, regardless of focus or cohort. By gender, the proportion of males and females in the sample was roughly equivalent, with slightly more boys overall (51%). With respect to age, all students were in Years 4-8 when their school commenced the project\(^3\), but the proportion of students in each year level is unequal due to the structure of schooling over these years. Many Year 6 students moved to intermediate schools (Years 7-8) or integrated intermediate/secondary schools (Years 7-13) at the end of the first year of the intervention. In addition, since students in Year 9 were not tracked, all students who were in Year 8 during the first year of the project were only followed for one year. There were also

\(^3\) Students start school at age 5 in NZ, so students in Year 4 are approximately 9 years old.
proportionally more contributing schools (Years 1-6) compared with intermediate (Years 7 and 8) and full primary (Years 1-8) schools.

By ethnicity, however, there were significant differences in the proportion of students from the major NZ ethnic groups, so these proportions are displayed in Table 1 comparing the LPDP sample against the asTTle normative sample since this is used as the comparison group. Unsurprisingly given the large number of students involved, these proportions differ significantly. However, the differences are unlikely to be large enough to have a substantive effect on the results, especially since the analyses presented throughout the thesis primarily focus on relative progress rather than absolute differences.

Table 1. Comparison of Ethnic Composition of the LPDP Sample Against asTTle Normative Sample.

<table>
<thead>
<tr>
<th></th>
<th>Reading</th>
<th></th>
<th>Writing</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LPDP</td>
<td>asTTle</td>
<td>LPDP</td>
<td>asTTle</td>
</tr>
<tr>
<td>NZ European</td>
<td>57%</td>
<td>62%</td>
<td>50%</td>
<td>56%</td>
</tr>
<tr>
<td>NZ Māori</td>
<td>22%</td>
<td>18%</td>
<td>22%</td>
<td>18%</td>
</tr>
<tr>
<td>Pasifika</td>
<td>8%</td>
<td>6%</td>
<td>14%</td>
<td>10%</td>
</tr>
<tr>
<td>Other</td>
<td>13%</td>
<td>13%</td>
<td>15%</td>
<td>16%</td>
</tr>
</tbody>
</table>

2.1.4. Tools used in this thesis

This section provides an overview of the tools used during LPDP, with a particular focus on assessment tools for teaching and learning (asTTle), since this is the assessment tool used in Chapters 3 to 5 to investigate the research questions of the thesis.

2.1.4.1. Assessment tools for teaching and learning (asTTle)

asTTle is a suite of assessment tools developed to assist teachers to evaluate student achievement and progress diagnostically. This tool has evolved over time, and several versions have been released, but the standardisation procedure and extensive curriculum mapping process and item validity verification has remained equivalent across versions. The cohorts
analysed in this thesis used version 4, which was the last non-electronic version. The current version is version 5, and is called e-asTTle (http://e-asttle.tki.org.nz/). The asTTle team no longer support version 4, and has removed this documentation from their website, so the details are largely adapted from the 2007 research milestone (Parr et al., 2007).

asTTle was developed as an educational resource that could be managed by teachers, referenced to the New Zealand curriculum, and attempts to identify the areas in which students need additional support, with respect to curriculum strands, achievement objectives, functions or levels. It assesses student literacy progress in reading or writing (or alternatively in mathematics if that is a focus) against the objectives established by the curriculum, which enables teachers to determine what to focus on in order to assist progress. asTTle has also published norms for each year level, based on the achievement of almost 100,000 students who provided data during the development phases and comprise the normative sample (Hattie et al., 2004). This sample was selected by choosing a nationally representative sample of schools with respect to location, size, type and decile. The asTTle tool was then further calibrated based on the data collected from this sample, alongside consideration of teacher feedback. As a result, the published norms are reflective of typical expected achievement and progress within New Zealand schools, providing a useful comparison group for the LPDP data.

An advantage of asTTle over other available tools is that there are comparatively few issues with ceiling effects, though on occasion there may be floor effects among very low achieving Year 4 students. asTTle reports also provide information about each of the areas of focus selected in the test, with a grade and numeric score provided for each. The grade scores are directly related to the curriculum levels developed by the New Zealand Ministry of Education. Within each curriculum level, dimensions are scored as basic, proficient or advanced. For example, a total score of 2b would reflect basic achievement at curriculum level 2. Figure 1 below shows the typical curriculum level expected for each year of schooling. The shaded area and overlap reflects the different ability levels of students within the same year level (Ministry of Education, n.d.). The asTTle grades were then mapped to a numeric score with a range of 100-1000, and
designed to have a standard deviation of 100 at each year level. Due to the complexities of analysing categorical grades, the analyses in this thesis use the numeric scores.

**Figure 1.** Typical pattern of curriculum level achievement by school year level (The New Zealand Curriculum Online, 2007).

**asTTle: Writing**

At the time of the research, asTTle writing was the only evidence-based nationally normed measure of writing achievement (Parr et al., 2007). This tool sees writing as motivated by communicative purpose or function, and therefore different scoring rubrics have been developed depending on which writing purpose is being assessed. The purposes identified were: persuade, explain, instruct, recount, narrate and describe. Students sitting the test are provided a number of writing prompts selected by the teacher that direct students toward writing for one of the six specific purposes. Scripts are scored against four “deep” and three “surface” dimensions of writing at each curriculum level. The deep features include rhetoric or audience; content inclusion; text structure/organisation; and language resources, while the surface features include grammar, spelling and punctuation.
Facilitators had considerable involvement with the use of asTTle with each cohort, so became increasingly conversant with the tool. As a result, they became aware of some inconsistencies in the language of the original indicators, so a new set of indicators was designed by a team funded by the Ministry of Education and used in the final cohort (2008–09). This has resulted in some minor differences in the way achievement was assessed between the two cohorts of data in this thesis. This has been kept in mind when conducting each of the analyses described in this thesis.

Analysis of the cross-sectional data from the normative sample indicated that there were differences in achievement patterns dependent on the specific writing purpose, and this is also noted in the asTTle technical manual. Unfortunately, the norms reported do not differentiate by purpose, meaning that if there were inconsistencies in the selection of writing purpose as the assessment focus across time points, some apparent gain (or decline) in scores would be expected simply due to the change in purpose. Facilitators and schools also argue that the purposes for writing differ with respect to the knowledge and skills required of students. These are important considerations for researchers investigating project effectiveness, since effectiveness is measured broadly by assessing the rate of progress of the students attending schools within the project, making changes in the assessment tools, tasks or foci a threat to the validity of such conclusions. However, these differences are typically small (10-15 points, or 0.1 to 0.15 standard deviations), and it is very unlikely that there were systematic shifts toward selection of an “easier” writing purpose, especially since there were no implications for teachers or schools with lower progress. In addition, the structure of the professional development provided by the LPDP involved facilitators working with schools and teachers closely, ensuring adequate consistency in the extent to which teachers/schools related the assessment task to the teaching and learning focus. Therefore, conclusions about the effectiveness of LPDP based on the series of analyses presented in this thesis are likely to remain valid even if there were some changes of the specific foci within certain schools over the course of the project. Any such differences are unlikely to affect the primary aims of the thesis, since all analyses are based on the same cohorts, meaning that any differences in the results are likely to be a result of the strengths and weaknesses of the type of analysis used, rather than specific features of the data.
Inter-rater reliability is a particular concern when assessing writing competency, especially since this was a new tool at the time of the first cohort in 2004. Given the degree of inexperience with the asTTle writing tool among schools in the initial cohort (not included in the analyses in this thesis), schools provided the scripts and results for a 10% sample of the students in their schools. At Time 1, all of these scripts were cross-scored by the research team to check agreement with the scoring by schools, and half were cross-scored at Time 2. This process was not considered necessary in the second and third cohorts (those presented in this thesis) since schools were more familiar with the tool, and facilitators were working closely with schools to ensure consistency. Research undertaken during the development of asTTle using the normative sample scoring demonstrated that teachers with minimal training were able to score scripts at a reasonable level of reliability, with inter-rater reliability around .7 for markers with no tool-specific training and .8 for those with a half-day of training (Glasswell & Brown, 2003; Brown, Glasswell & Harland, 2004). Glasswell and Brown state that reliability of .8 is sufficient to “support high-stakes interpretations based on the obtained scores for writing” (2003, p. 1). Given the facilitation structure of LPDP, the reliability of asTTle writing scores within the dataset being used in this thesis is likely to reflect that of trained markers, especially since facilitators were heavily involved in moderation of the first set of scripts in each cohort to ensure that reliability was as high as possible. Examples of pre-scored texts, along with scoring tips, are available for each purpose from the e-asTTle website (http://e-asttle.tki.org.nz/) under teacher resources.

**asTTle: Reading**

Reading ability is assessed and reported against curriculum objectives and processes. A reading test bank was developed, with all questions having been categorised by teacher panels according to the relevant objectives and dimensions of the NZ Curriculum. The panels made decisions about items with respect to three processes: exploring the English language; critical thinking about language and meaning; and processing information to identify, understand, store, organise, retrieve, combine, and communicate the information (Parr et al., 2007). The development of these curriculum-indexed test items related to the idea of surface and deep
cognitive understandings, based on the Structure of the Observed Learning (SOLO) taxonomy, whereby students are assessed in terms of the quality and complexity of their work, rather than just how many isolated pieces they get right (Biggs & Collis, 1982). Items that measure surface understanding need only knowledge of one fact or idea given in the question, or knowledge of more than one idea, but each used separately and without integration. Deep processes are measured via evidence of a shift in the quality of thinking, where individual pieces of knowledge are integrated. The highest level of this would involve a student extending this beyond the information given to extract, for example, a general rule. Deep processing in reading relates to finding information, knowledge, understanding, connections and inference, with individual items sometimes assessing more than one of these.

When setting up an asTTle reading comprehension test, teachers are able to select the difficulty level with respect to the curriculum level, and decide the proportion of questions to come from specific curriculum levels, with the asTTle programme selecting some additional items outside the levels chosen by the teacher. In addition, teachers select up to three processes to test, with asTTle again selecting some items from outside the specific focus of the test. This adaptability means that there are several thousand possible test variations, but the scores from each test are comparable since items were developed using item response theory (IRT) to place items and students on a common scale. IRT models emphasise the responses to each item, rather than simply looking at the overall test results, and theorises that the probability of a correct response is a mathematically measurable function of the respondent, and the item (Baker, 2001). In other words, IRT models both the difficulty of the items, as well as the ability of the students, on a common, standardised scale. This allows tests made up of different items to be directly comparable since the difficulty of each item can also be compared. IRT is also an essential component of the adaptive testing inherent to the current, electronic form of asTTle. This ensures that generally scores on the asTTle scale are comparable irrespective of the level of the test.

However, it is still important to identify the correct test level so that the results reflect students’ ability rather than the difficulty of the items, by including items that provide students the opportunity to show what they know. This is partially compensated for by the automatic addition
of some items outside the test levels selected, but not completely. Maximum consistency in the way the test levels were selected was considered essential since the project focused on reducing disparities via heightened rates of progress, and an inappropriate level setting could introduce a confounding effect. This is emphasised by an analysis with 96 students who were tested at least twice at baseline. This was generally because the teacher thought that the student's test result was not reflective of his/her ability. It is unclear which test was given first, but the tests were all administered at baseline, within the same term of schooling. On average, students scored more than 70 points higher when tested using a higher level test. Three students were tested three times within one time point; in all cases, these students scored higher, the higher the level of the test, making a difference of more than 200 points for two of the students (150 for the third). This shows how imperative it is that teachers choose a test that is appropriate to the ability of individual students.

For this reason, the research team constructed the asTTle reading tests during the first cohort. During the second and third cohort, the test levels were set by teachers with the help of facilitators since facilitators were highly knowledgeable about asTTle by this time, and able to assist with setting appropriate test levels. The facilitation structure of LPDP is considered likely to have generally ensured that test levels were appropriate since teachers were provided with significant scaffolding when setting test levels at baseline, but it does remain a possibility that at least some of the reading progress noted in LPDP could be explained by teachers becoming increasingly capable of assessing their students' abilities and selecting tests that allow them to demonstrate these abilities.

The tests set by the research team during the first cohort are described below so that the reader has a clear picture of the test structure. All asTTle reading tests are set in a similar manner, with the teacher or test administrator making the decisions described above. Since each item may measure several aspects of reading (this is quantified using the IRT analyses), the number of items in a test is considerably less than the number calculated by simple addition of the items in the test structures described below.
The level 2 reading test was designed for administration to Year 4 students, and was mainly comprised of items at level 2, plus a few at level 3. The resulting test had six items at 2B; seven at 2P, ten at 2A, three at 3B, one at 3P, zero at 3A, one at 4B and one at 4P. The processes chosen as foci included finding information (some items), knowledge (some) and understanding (most). This resulted in 11 items related to finding information, 10 knowledge items and 29 understanding items. The automatic addition by asTTle resulted in the inclusion of three connections items and five items dealing with surface features. There were 17 items related to surface features of reading and 18 deep feature items.

The level 3 reading test was administered to students in Years 5 and 6, and included some items at level 2 and most at level 3 with a few at Level 4. This resulted in a test with two items at 2B, six at 2P, five at 2A, eight at 3B, five at 3P, three at 3A, five at 4B and one at 4P. The foci of the level 3 test included knowledge (13 items), understanding (29 items) and inference (30 items). However, although these were the processes selected by the research team, the automatic addition by asTTle resulted in nine items related to finding information four related to connections. There were 15 surface feature and 22 deep feature items.

The test designed for intermediate school level students (Years 7 and 8) was comprised of a few items at level 2 and 4, with most items at level 3. The result was four items at 2B, one at 2P, zero at 2A, eleven at 3B, eight at 3P, five at 3A, and five at 4B. The research team chose understanding (16 items), connections (11 items) and inference (23 items) as the processes to focus on in this test. The items selected by the asTTle programme related to dealing with finding information (14 items), knowledge (five items) and surface features (one item). There were 12 surface feature and 23 deep feature items.

2.1.4.2. LPDP data not used in this thesis

It is important to note that LPDP collected a range of other data in addition to the student achievement data collected using asTTle, that are not used for the analyses presented throughout this thesis. Some of these data also related to student achievement, while the
remainder examined aspects of teacher/leader practice and student beliefs using a mix of surveys and interviews. These other data were excluded for various practical reasons. The student achievement data collected using other tools were typically constrained by the distributional properties of the tools, while the survey and interview data did not directly relate to the primary aim of the thesis. While it is likely that these other data would be useful to help explain the project effectiveness, their inclusion would shift the focus away from the analysis and back to the substantive content, which is not the focus of this thesis.

Achievement among younger students (Year 1) was measured using the Observational Survey of Early Literacy Achievement (Clay, 2006), which tests basic literacy skills. The Observational Survey is a useful diagnostic tool but has a very significant ceiling effect so is of limited utility for measuring progress among students above the median. Reading comprehension of Year 3 students was tested using the Supplementary Test of Achievement in Reading (STAR; Elley, 2003). STAR reports achievement using stanines, which divides a theoretical normal distribution of results into nine categories resulting in a mean stanine of 5, with a standard deviation of 2. Since there are only nine possible categories, the variance is substantially reduced limiting the flexibility of analysis choice, so results from this test were not examined in this thesis.

As noted in the previous section, in the final cohort (2008-2009), schools were given the option of using the Progressive Achievement Test (PAT; Reid & Elley, 1991) to measure achievement and progress among students in Years 4-8 as an alternative to asTTle. As with STAR, PAT reports stanines, limiting the range of analyses that could be appropriately applied (subsequently, both tools have been updated to include a more continuous ‘scale score’). More importantly, the majority of schools opted to continue using the asTTle test, so the asTTle data remained the focus of the analyses conducted in this thesis.

Survey and interview data were collected from a subset of schools (case study schools). For example, one questionnaire sought information regarding facilitator and teacher perception of lesson quality following a lesson observed by a facilitator was also collected, while another collected a range of information about leadership, school focus and climate, providing useful descriptive information about the school and its practices. Students were also asked about their
learning views, in order to find out what they felt was important in becoming successful readers/writers. Additional measures were also collected, though the above are the most likely candidates with respect to value in explaining differences in progress. The explanatory value of these measures will be examined in future publications.

2.1.5. Limitations of the LPDP dataset

As with any large-scale, school-based educational data, there are a number of limitations that need to be discussed. Generally it was possible to correct and control for these limitations since considerable effort was expended implementing data checking and validation procedures. Some limitations arose from the specific tool used; these have been discussed in the section describing asTTle. Remaining limitations generally related to discrepancies between intent and implementation, in that the procedures and processes envisioned by the LPDP team with respect to data collection were not always identical to those delivered (or indeed deliverable) by facilitators. This was likely exacerbated by insufficient data literacy – future interventions may do well to place more emphasis on the importance of accurate data collection and collation from the outset, since data provide an essential contribution toward evidence of project effectiveness.

The first limitation relates to attrition. This is a major concern with longitudinal data (Twisk & de Vente, 2002), since non-random attrition produces biased results with many of the commonly applied methods of analysis, meaning that one cannot simple ignore the missing results. There was loss of student data due to absence caused by illness or otherwise, as well as through change of schools. As indicated earlier in this section, most of the attrition occurred due to students changing schools at the end of Year 6, since most primary schools in New Zealand only cater for students up to Year 6. Overall, longitudinal data were obtained for 75% of students in the 2008-2009 cohorts, and for 60-65% of students (writing and reading, respectively) in the 2006-2007 cohorts. Typically, listwise deletion of students with missing data is inadvisable, with the potential for substantially biased results when data are not missing at random. Imputation was not appropriate in this case since the values that were missing were the dependent variables (i.e.,
the achievement score), rather than the independent variables (such as gender and ethnicity). As a result, using imputed scores would merely add noise to the analyses, without contributing any additional information (von Hippel, 2007). There are procedures that are appropriate for imputing missing dependent variables when these measures are repeated, but these are more appropriate when there are several observations. In order to mitigate the effect of attrition, and include as many students as possible, procedures requiring students to be present at all three time points have been avoided in this thesis. All analyses presented include every student for whom two observations were available, while the analyses in Chapter 5 present procedures with which missing data have less effect.

An additional difficulty related to data collection and longitudinal tracking. The use of automatically generated asTTle files reduced the likelihood of errors caused by inconsistencies across schools, but there were still problems with some of the optional or changeable fields. In particular, this related to the use of student IDs, or differing naming conventions. Some schools used an ID that changed at the beginning of each year, while others did not provide any ID at all. The asTTle file also includes the student names (data were anonymised after aggregation), but in approximately 5% of cases, the reported name changed over the course of the project. These apparent name changes were generally caused by errors in spelling, or the use of a nickname at one time point and not another. As a result, there were several hundred cases that would not have been matched had we simply relied on matching identical cases (i.e., the same student ID, or name). Due to the large number of students involved there were also several cases where students from different schools had identical names. As a result, an approximate string matching algorithm was developed that incorporated information from all available demographic (gender, year level, ethnicity) and locational information (school) as well as the more stable aspects of students’ names (surnames, and first initial) to indicate the likelihood that the results were from the same student. Instances where there was a complete match were merged automatically, while those that produced a high likelihood (above 80%) were checked manually before making a decision about whether to merge the cases. Only cases where it was clear that the students were
identical were matched, since incorrect matching was likely to create more bias than leaving a small number of unclear cases unmatched.

There is also a limitation inherent with the categorisation of particular subgroups. For example, NZ European, Māori and Pasifika are treated as distinct subgroups, and are categorised by whichever of these groups they feel they identify with most strongly. However, there is frequently overlap, and the individuals in such categories are not necessarily cohesive. This is clear when one looks at the pattern of achievement of these subgroups. Typically the focus is on the difference between subgroups, but it is important to note that there is more difference within these groups than between them due to considerable distributional overlap.

The asTTle scale introduces some difficulty when making comparisons across schools in both reading and writing using simple averages since there are differing numbers of students in each year level, with different schools having differing proportions in each of the year levels. As the national norms for year levels have varied markedly in versions of asTTle used, these differing proportions need to be taken into account in calculating group or school-level average measures of progress.

Finally, generalizability is also a limitation. Although the analysis of methodological choice undertaken in the thesis is highly generalizable and likely to be useful to other researchers, the findings explaining progress must be viewed contextually – that is, within the specific intervention in which the data were collected. The majority of students attending schools taking part were assessed using the asTTle tool, so this provides a good picture of the achievement and progress profile within LPDP. However, the degree to which factors such as ethnicity and gender relate to differences in achievement, is likely to have been altered by involvement in the project, so only the baseline measurements can be expected to have results reflective of the wider population (i.e., before the project was able to have an effect). In addition, since the schools are self-selecting, there is a bias toward inclusion of schools (and students) with greater perceived (and actual) levels of need, resulting in baseline results considerably below the national average. Still, the project is very large, especially for the New Zealand context, so represents a considerable proportion of the NZ schooling context albeit with some bias toward students with greatest needs.
These limitations reflect the realities of data collected from naturalistic educational settings. With such a large project, it would not be possible to ensure that student achievement data were obtained for all students at all three time points over a two-year period; some students will change schools, and schools cannot be mandated to remain within a project if they choose to opt out, due to the degree of autonomy allowed to individual schools within the New Zealand educational system. Awareness of these limitations is essential as this awareness must be incorporated into the analyses conducted, as well as the conclusions made about the research.

2.1.6. Assumptions made in this thesis

This thesis focuses on quantitative methodologies that are frequently reported in the literature to explore various explanations for differential progress among students of ostensibly similar ability (estimated by initial achievement), applying these methods to a specific intervention (LPDP). The fundamental assumption of this thesis is that asTTle is an appropriate test of initial ability and progress and that, on average, the scores will reflect the true nature of the ability of a group of students. As such, it is also assumed that the overall effect of various potential extraneous influences on scores, such as teacher expectations, illness and noise, will be relatively evenly distributed across all students, and so will not introduce bias for specific analyses. In addition, since asTTle scores needed to be ‘standardised’ across year levels, it is essential that the published norms provide at least a reasonable estimate of the typical achievement among students for each year level. If there is a large departure in the appropriateness of the norm for a particular year level, the uneven weighting of students in each year could unduly affect the results. The statistical phenomenon of regression to the mean is considered in the regression chapter (Chapter 4) since this is a definite, non-random bias that can be estimated.
2.2. Chapter Summary

This chapter described the structure of LPDP and the tools used during the project, with a particular focus on the data used in the analyses presented throughout this thesis. While the primary aim of this work is to emphasise the importance of appropriate use of quantitative methodology, the analyses presented throughout this thesis are also of use to researchers interested in professional development interventions. The results described throughout provide strong evidence of the strengths and weaknesses of the procedures followed during LPDP and it is anticipated that these results can be combined with the information and findings presented in the LPDP milestones and used to guide and improve future professional development work.
CHAPTER 3. EFFECT SIZES

The previous chapter introduced the Literacy Professional Development Project, from which the data analysed throughout this thesis were drawn. The current chapter makes use of the LPDP data to investigate the progress of students in this project using alternative measures of effect size to address the thesis research questions. As noted in Chapter 1, since effect sizes are considered accessible to the wider educational community, the reporting of effect sizes has been virtually mandated by the APA since 2001, and is a requirement or strongly suggested by a number of large research organisations and major journals (AERA, 2006; APA, 2001; 2010; Capraro & Capraro, 2003; JEP, 2003; La Greca, 2005; Peng et al., 2013), but several effect size metrics are available and even the more popular measures are not always directly comparable. Therefore, the extent to which conclusions differ as a result of the use of these different effect size measures provides a preliminary insight into the primary aim of the thesis; that is, the extent to which the choice of analysis really matters. In addition, instances where results are relatively stable across the different effect size metrics begin to build a picture about whether LPDP can be deemed to have produced substantive shifts in student achievement.

The first section of this chapter provides a general review of the concepts relating to the standardised mean difference (SMD) effect size, which is currently the most popular effect size metric. However, there are substantive differences in the literature about precisely how the SMD should be calculated, as well as the extent to which issues such as dependency and the distributional properties of the data should be taken into account, so these issues are discussed in subsections 3.1.1 and 3.1.2. The following section (3.2) introduces alternative effect size metrics and presents some of the arguments in the literature about the use of these metrics. The two alternatives chosen as comparative methods to be presented in the analysis section are introduced in greater detail in subsections 3.2.1 (odds ratios) and 3.2.2 (Cliff’s delta). The analysis section (3.3) presents the results of the effect size analyses of progress, applying the methods introduced in the earlier sections of the chapter to the LPDP data. Odds ratios are presented in a separate subsection (3.3.1), while the SMD and Cliff’s delta are presented in the
same subsection so that the results can be directly compared. The final section of the chapter summarises the major findings of the chapter, with a particular focus on differences in conclusions about project effectiveness resulting from the different methods; thereby relating the results to the primary aim of the thesis.

3.1. Standardised mean difference

A major driver of effect size usage has been the massive increase in popularity of meta-analysis since the term was first coined, and the methodology formalised by Glass more than 30 years ago (Glass, 2000; Ahn et al., 2012). Glass (1976) first defined meta-analysis as “the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings” (p. 3). The need for formalised meta-analysis was driven by a tendency at this time to focus solely on study design, with considerable bias toward only acceptance of randomised controlled trials published in peer-reviewed journals – Glass even goes so far as to argue that all too frequently this process resulted in the included studies “being one’s own work or that of one’s students or friends” (1976, p. 4), and instead argued for a far more inclusive meta-analytic process of aggregation. As such, meta-analysis is intended to afford the benefits of greater statistical power due to increased sample size, and therefore increased generalizability, providing better evidence of the underlying “true” effect (Cohn & Becker, 2003). In order to integrate the results from often quite disparate research methodologies, some formal measure by which the results of a series of studies aggregated according to some common metric must be derived.

Glass (1976) used the still popular effect size metric of standardised mean difference (SMD) in this first meta-analytic illustration, which likely explains at least some of the popularity of such measures. Indeed, he notes that despite the notion of effect sizes having been around for decades prior to his original paper on meta-analysis, his main contribution to meta-analysis has repeatedly been cited as having provided the idea to divide mean differences by standard deviations as a form of research synthesis (Glass, 2000). It is interesting to note that Glass
describes considerable discomfort with this attribution, lamenting the fact that the research community appears to have ignored the considerable early work he conducted with respect to alternative effect size measures.

The effect size movement garnered considerable additional momentum after Cohen’s seminal work (1988, cited more than 50,000 times; 1994) denouncing the usage of simple significance testing even for individual studies, based on the assertion that null hypothesis significance testing (NHST) only provides information about the reliability of a relationship, but nothing about the strength, or “meaningfulness” of the effect being measured (e.g., Rosenthal, 1994; Trusty, Thompson, & Petrocelli, 2004). However, despite an absolute wealth of literature discussing the various effect sizes measures available (e.g., Coe, 2002; Cohen, 1988; Glass, McGaw, & Smith, 1981; Grissom & Kim, 2005; Fleiss, 1994; Hedges & Olkin, 1985; McGaw & Glass, 1980; Morris & DeShon, 2003; Rosenthal 1991; 1994; Rosenthal, Rosnow, & Rubin, 2000) there is considerably less guidance around appropriate effect size usage (Baguly, 2009; Coe, 2002; Durlak, 2009), with many researchers either calculating and reporting effect size measures that are inappropriate given certain properties of their data, or metaphorically throwing the baby out with the bathwater and opting to report effect sizes in isolation, neglecting to use NHST to ascertain whether the relationship is actually reliable (Cortina & Landis, 2011; Peng et al., 2013).

SMD effect sizes have remained overwhelmingly the most popular metric in the social sciences (Bloom, Hill, Black, & Lipsey, 2008; Peng et al., 2013). Indeed, the results of the almost 150,000 studies synthesised in Visible Learning (Hattie, 2009) were aggregated in this way. A number of very minor adjustments have been suggested by various researchers whose names have frequently been appended to the procedure. Glass’ $g$ initially measured effect size by simply dividing the group mean difference by the standard deviation of the control group (Glass, 1976; 1977). This was noted to be problematic since it is not always clear which group is the control in educational interventions, and therefore, which standard deviation to use (Coe, 2002). Cohen’s $d$ also defined effect size as the difference in group means divided by the standard deviation, but did not explicitly state which standard deviation, instead stating that the standard deviations of
both groups are assumed to be equal (Cohen, 1969). This was later formalised as the pooled standard deviation, and was calculated using the following formula:

\[
s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2}}
\]  

(1)

This proposes that \( s \), the standard deviation described by Cohen, be calculated by pooling the standard deviations of both groups, while adjusting for differences in the sample size of each group. However, this calculation was shown to be positively biased since the equation would result in an estimated pooled standard deviation smaller than the individual standard deviations in instances where the variance was identical in both groups, resulting in an apparently larger effect size estimate (Hedges, 1981). Hedges (1981) proposed a further minor adjustment to the effect size calculation in which the degrees of freedom of the pooled standard deviation is adjusted by subtracting two from the denominator to compensate for the degrees of freedom adjustment in the numerator, as shown in the equation below:

\[
s' = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}
\]  

(2)

However, while this resulted in a reduction of the bias described above, a small bias remained (Hedges & Olkin, 1985). Hedges and Olkin (1985) proposed a correction factor that provides an unbiased estimate of the effect size using the gamma function. Since the gamma function is difficult to calculate the approximation below is typically used. This equation approximates the gamma function correction \( J \) very closely, with an error of 0.0003 with 10 degrees of freedom (this error declines as \( n \) increases).

\[
J \approx 1 - \frac{3}{4df - 1}
\]  

(3)

In order to correct the positive bias shown in Cohen’s \( d \), the effect size estimate derived is multiplied by the correction factor, \( J \). The \( df \) in the approximation of the correction factor is defined as the total sample size minus two (i.e., \( n_1 + n_2 - 2 \)). The resultant effect size is referred to as
Hedge’s $g$, which provides an unbiased estimate when normality and independence assumptions are met, but remains susceptible to violations of normality.

### 3.1.1. Dealing with dependency

Effect sizes were initially conceptualised as a way to assess and compare the relative magnitude or meaningfulness of different effects. This initial conceptualisation focused on measuring effects within an experimental-control research design, with little discussion of how to measure the effect size in repeated measures designs. Since educational research frequently uses a longitudinal design, for example, tracking student achievement over time, it is important to consider how to deal with the dependency inherent in longitudinal data, since this is a typical assumption of statistics. Unfortunately, there is no clear consensus in the literature about how dependency should be treated with an array of different assertions about the most appropriate procedure, often resulting in very large differences (frequently two or three orders of magnitude) in the effect size estimate depending on which procedure researchers opt to follow (Dunlap, Cortina, Vaslow & Burke, 1996). As with the various bias adjustments discussed in the previous section, these recommendations all centre on adjustments to the standardiser (i.e., the standard deviation used in the denominator of the effect size equation) since this is how the mean differences are scaled in an attempt to produce a comparable metric, but the implications are far greater. Researchers frequently fail to indicate whether dependency has been accounted for, or even which standardisation procedure has been used, so there is a considerable threat to the validity of meta-analyses that combine effect sizes derived from different research designs since the various approaches have the potential to produce such large differences in the apparent effect size magnitude (Ahn et al., 2013; Morris & DeShon, 2002).

Current approaches to effect size calculation for repeated measures studies follow one of two main “families”. The most common approach, herein referred to as the "raw score" approach, provides an effect size estimate very similar (or identical) to that derived using the equation for independent effect sizes. One such approach was proposed by Becker (1988), who argued for
standardisation using the pre-test standard deviation, since variance at pre-test is likely to be unaffected by the intervention and therefore more stable – mirroring the argument for Glass’ $g$ (standardisation using the control group standard deviation). Others advocate an approach more in keeping with Cohen’s $d$, proposing that the raw pre- and post-test standard deviations be pooled, arguing that each raw estimate of standard deviation can be considered independent estimates of the population standard deviation, and thereby increasing the reliability of the estimate (Dunlap et al., 1996; Hunter & Schmidt, 1990; Taylor & White, 1992). This assertion is argued to be problematic since the increased reliability cannot be measured, making the standardisation inconsistent across studies when variance differs over time (Morris & DeShon, 2002).

The second main approach uses the standard deviation of the change scores, arguing that the raw score approach ignores the dependency in the data (Gibbons, Hedeker & Davis, 1993; Johnson & Eagly, 2000; Rosenthal, 1991). Further, these researchers note that since standard deviations are often very small at pre-test due to various selection conditions, for example, a study targeting students with very low initial achievement, it is generally unlikely that normality assumptions will be met when using the raw score procedures. Conversely, change scores are much more likely to be normally distributed; meaning that standardisation using the change score standard deviation is more widely applicable.

Unfortunately, there remains considerable debate about the most appropriate approach. Dunlap and colleagues (1996) argue convincingly that dependency is something of a red herring with respect to effect sizes, asserting that while there is increased reliability in the statistical estimates derived using paired designs, there is no basis for adjusting the magnitude of the effect size. In other words, while statistical significance is affected, practical significance should not be. Morris and DeShon (2002) insist that both approaches are valid, but appropriate for different research questions. When the researcher is interested in differences between conditions that arise over time (e.g., when there is an experimental and control group), the raw score approach is argued to be more appropriate, while research investigating individual progress is better served by the change score approach (Morris & DeShon, 2002).
3.1.2. Distributional considerations

Although there is considerable discussion in the literature about the instability and bias associated with effect sizes based on standardised mean differences when normality assumptions are not met (i.e., Algina, Keselman & Penfield, 2005; Hedges & Olkin, 1985; Maxwell, Camp & Arvey, 1981; Wilcox, 1995), most researchers continue to use variants of the SMD – while frequently failing to indicate which variant of the SMD effect size they have presented (Peng et al., 2013). The popularity of SMD effect size measures is arguably a result of ease of calculation rather than any particular suitability for the data being analysed, and there is an increasing body of literature questioning the continued use of these measures without consideration of the distributional properties of the data (i.e., Barnett, van der Pols & Dobson, 2005; Coe, 2002; Cortina & Landis, 2011; Fern & Monroe, 1996; Hess & Kromrey, 2004; Hogarty & Kromrey, 2001; Ledesma, Macbeth & Cortada de Kohan, 2009; Leech & Onwuegbuzie, 2002; Romano, Kromrey, Corragio & Skowronek, 2006).

This is further exacerbated by underlying differences in the distributional and contextual properties of the various studies being compared (Hogarty & Kromrey, 2001; Lipsey et al., 2012; Trusty et al., 2004). Interpretation of effect sizes must take into account features such as study duration, cost and additivity (whether one effect adds to another), as well as features of the assessment, such as reliability, floor/ceiling effects, whether the effect size has been derived from a continuous scale, or reflects shifts over and above maturational progress (as is the case with many standardised tests). These considerations do not invalidate the use of effect sizes necessarily, but they frequently jeopardise the validity of the conclusions made. The most common concern about SMD effect sizes is that they rely heavily on normality assumptions, which are relatively rarely met in educational contexts (e.g., Micceri, 1989; Hess & Kromrey, 2004; Ledesma et al., 2009; Leech & Onwuegbuzie, 2002; Romano et al., 2006), meaning that means and standard deviations are frequently a poor representation of the distributional properties of the data. Regression to the mean compounds these problems and, given that many interventions focus especially on lower achieving students, the apparent effect size in these cases is, by definition, larger simply due to regression to the mean (Barnett et al., 2005).
For example, Anand and Bennie (2005) reported on the effect of Reading Recovery in New Zealand on student achievement. Almost 11,000 students were followed and mean gain scores reported as evidence of the success of the intervention. Effect sizes were not reported in this article, but were calculated and reported subsequently as part of a Best Evidence Synthesis (Timperley et al., 2007). This is common practice in meta-analysis, but is not always appropriate. In this instance, the entry data were severely skewed (by definition, those entering Reading Recovery must have very low initial achievement). The resultant low variance in starting scores (all started low), combined with the statistical phenomenon of regression to the mean, results in a substantially inflated effect size. A similar article in the United States investigating the efficacy of Reading Recovery against other literacy programs, such as small group tutorials and the Success for All project (SFA; mentioned in Chapter 1) also used effect sizes, calculated using the same formulae (Slavin, Lake, Davis, & Madden, 2011). Selection criteria were far stricter, only including studies with a control group, and the distributional differences at pre-test were less than half a standard deviation. However, the relative starting points of students included in the studies compared with those in the general population is not considered, meaning that regression to the mean will have differing impact on the effect sizes, dependent on the relationship between pre and post scores, as well as the degree of extremity of the student scores included. As the focus of the article was on children with reading difficulties, this is particularly pertinent due to the differing definitions included (some studies included all children in the lowest third, while others only included those with diagnosed reading “disabilities”). The phenomenon of regression to the mean is considered further using the LPDP data later in the analysis section of this chapter.

Timperley and colleagues (2007) identify low starting scores as an important contextualisation issue, along with several other pertinent considerations such as differences in target versus test (i.e., whether the intervention specifically targets aspects of learning tested for, or focuses on more general development) and differences in typical effect sizes observed depending on the particular methodology used (e.g., those with a control group typically have lower effect sizes since the control group exhibits maturational growth). Baguly (2009) argues that SMD effect sizes are especially susceptible to differences in distributional position; that is, whether individuals
score low or high on the variable of interest since SMD effects only measure the effect at the centre of the distribution, as well as differences in study design, choice of test measurement and so on, since all of these factors have a considerable effect on variance. These issues are common to all forms of meta-analysis and while they can never be completely overcome, there is substantial room for improvement in the general literature as evidenced by the large number of articles discussing the potential disparities in effect size meaning across different studies (e.g., Fern & Monroe, 1996; Hogarty & Kromrey, 2001; Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006; McGaw & Glass, 1980; Morris & DeShon, 2003; Romano et al., 2006; Sánchez-Meca, Marín-Martínez, & Chacón-Moscoco, 2003).

3.2. Alternative effect sizes

Alternative effect sizes include the so-called $r$-family, which are also standardised measures, but are based on the amount of variance explained by a relationship. The most common of these is Pearson’s correlation coefficient, or $r$, which conceptualises the strength of the relationship as being analogous to the size of the effect (Rosenthal, 1994). However, these measures have the same issues around vulnerability to violations of normality, design-related differences in variance and reliability (Baguly, 2009), so are not discussed in detail. Other alternatives argued to be more widely suitable, include odds-ratios for reporting categorical effects (Durlak, 2009), or non-standardised measures such as simple mean group differences for continuous data, since these measures are not affected by distributional variance (Baguly, 2009; Durlak, 2009), or non-parametric alternatives, such as Cliff’s delta (Cliff, 1993; 1996). The non-standardised group difference measure advocated by Baguly (2009) still relies on means as a suitable reflection of the distribution, which results in the measure remaining extremely susceptible to violations of normality. In addition, Durlak (2009) notes that such differences are very difficult to compare in a meaningful way since the measures and assessments used across studies frequently have markedly different properties. Since non-standardised effect sizes present a somewhat flawed alternative, and can be easily derived from the standardised measures, this approach has not
been used to analyse the LPDP data in this thesis. As already noted, standardisation does not entirely mitigate the issue of comparability, but it does provide a more comparable metric than leaving it on the original measurement scale.

### 3.2.1. Odds ratios

Odds ratios (OR) are useful for comparing the relative odds of an outcome occurring for one group compared with another. An odds ratio is calculated according to the formula below:

\[
OR = \frac{a/b}{c/d}
\]  

(4)

Where \(a\) and \(c\) are the number of outcomes of interest, in the intervention (or group of interest) and the control (or comparison/reference) groups, respectively. For example, in order to investigate the relative odds of girls passing a particular assessment compared with boys, girls could be designated as the groups of interest, and boys, the comparison. If there were 60 girls that passed the assessment compared with 50 boys, \(a\) would be 60 and \(c\) would be 50. The letters \(b\) and \(d\) in the equation refer to the number of instances where the reverse outcome, again in the intervention and comparison groups respectively (Durlak, 2009). If there were 100 students of each gender, \(b\) would be \(100 - 60 = 40\), while \(d\) would be \(100 - 50 = 50\). Therefore, the odds of passing would be 60/40 for girls; that is, girls are 1.5 times more likely to pass than to fail, while the odds of passing for boys would be 50/50; that is, boys are equally (1 times) likely to pass as fail. The overall odds of passing for girls relative to boys would be as follows:

\[
OR = \frac{60/40}{50/50} = 1.5
\]  

(5)

Therefore, the odds of a girl passing this example test are 1.5 times greater than for boys. Another useful property of the odds ratio is that ORs are inversely proportional. That is, an OR of 1.5 is symmetrical to \(1 / 1.5 = 2 / 3 = .67\), so the odds could be restated with boys as the ‘interest’ group, in which the odds of boys passing the test was two-thirds that for girls. Odds ratios are useful in that the underlying distributions are effectively unimportant (Fleiss, 1994; McGrath &
Meyer, 2006), but limited in the sense that they require data to be categorised in a binary manner (i.e., exhibited outcome of interest or not). An example of how odds ratios can be used is presented using students in the lowest 20% of the LPDP sample in the analysis section of this chapter, but is not developed across all analyses due to the limited range of applications within the educational sector compared to the other effect size options presented.

3.2.2. Cliff’s delta

Arguably the most convincing alternative to the standardised mean difference family of effect sizes is the use of a non-parametric alternative such as Cliff’s delta ($\delta$). Non-parametric effect sizes do not make assumptions about underlying distributional properties, meaning that these measures are comparatively unaffected by variance heterogeneity and violations of normality, producing substantially more robust and reliable metrics, which is especially useful when attempting to compare the magnitude of effects across studies (e.g., Grissom & Kim, 2001; Hess & Kromrey, 2004; Ledesma et al., 2009; Leech & Onwuegbuzie, 2002; Romano et al., 2006).

Lack of normality is a pervasive issue, with Micceri (1989) arguing that the majority of data collected within the social sciences do not meet the basic normality assumptions necessary for accurate parametric analysis – going so far as to liken the likelihood of data collected in applied settings being normal as equally likely as an encounter with a unicorn. This is even more problematic given that separate reviews of the literature by Breckler (1990) and Osborne, Christiansen and Gunter (2001) indicated that that fewer than 10% of researchers actually even consider whether normality assumptions are met. Since SMD effect sizes are measured in standard deviation units calculated using only the means and standard deviations of the two groups being compared, they are extremely vulnerable to violations of normality and not an appropriate quantification of the effect (Hogarty & Kromrey, 2001). However, even in instances where researchers have acknowledged that their data are not normally distributed and have appropriately employed the use of non-parametric statistics to test the null hypothesis, many persist in reporting SMD or other parametric effect sizes (Leech & Onwuegbuzie, 2002). The
most obvious reason for this is the paucity of information about how to select and calculate a non-parametric effect size, coupled with the inability of major statistical packages to compute it for researchers (Hess & Kromrey, 2004; Ledesma et al., 2009; Leech & Onwuegbuzie, 2002; Romano et al., 2006).

Cliff’s delta is a robust and intuitive alternative to Cohen’s d especially useful in situations where data are either non-normal, or are ordinal and therefore have reduced variance (e.g., likert scale responses from survey data), though there is no down side (other than being slightly more complex to calculate) to using the metric even when normality assumptions are met (Cliff, 1993; 1996). Cliff’s delta was originally conceptualised as a dominance statistic (Cliff, 1993) and is obtained by calculating the non-overlapping area of the two distributions at the case level. The statistic ranges from −1 to +1, where either extreme indicates no overlap between the two distributions and 0 indicates complete overlap. A particular advantage of δ is that it measures the overall magnitude of an effect across the entire distributions of both groups, rather than simply the size of the effect for those in the centre of the distribution which is a common criticism of SMD effect sizes (Grissom & Kim, 2001). There are a number of articles that advocate the use of δ and emphasise its robustness in situations where parametric assumptions are not met since the metric is robust under violations of normality, even when the groups have uneven numbers and / or variance (e.g., Fern & Monroe, 1996; Ledesma et al., 2009; Romano et al., 2006; Hess & Kromrey, 2004), but few give advice on how to actually calculate and interpret the statistic.

As might be clear, measuring the proportion of distributional non-overlap, while conceptually simple, is a relatively involved task requiring matrix statistics. Effectively, a matrix is constructed to determine the proportion of instances where scores from one group are larger than the scores from another; in other words, the ratio of “dominance” of one group over the other as shown in the calculation is formalised in the equation below:

\[ \delta = \frac{\#(x_{i1} > x_{j2}) - \#(x_{i1} < x_{j2})}{n_1n_2} \]  

(6)
This equation states that the estimate of delta is derived by subtracting the number of instances of dominance in the comparison group from the instances of dominance in the group of interest, with the result divided by the total number of comparisons, calculated by multiplying the sample size of each group. All possible comparisons are made, meaning that the entire distribution is represented by the statistic – which does result in considerable computational load when working with large samples. The choice of which group is subtracted from the other is only important in that it will affect whether the statistic is positive or negative (and therefore conclusions must take this into account), but will not affect the degree of non-overlap. The magnitude of the effect size described by δ is easily quantified with respect to Cohen’s d since Cohen (1988) also provided an interpretation of SMD effect sizes in terms of distributional overlap. Therefore, the degree of overlap can be determined by subtracting Cliff’s δ from 1 (since δ describes the non-overlap), then converted into the unbiased SMD that would result if the distributions were normally distributed. Since this “bridge” is based on normal distributions, the degree of discrepancy between the SMD estimate derived using Cliff’s delta and the SMD calculated in the usual manner, demonstrates the degree of bias that arises from the use of a parametric effect size for non-normal data. This is demonstrated in the analysis section of this chapter.

To assist with interpretation of δ, Figure 2 shows the degree of overlap associated with a moderate Cohen’s d of 0.5. The overlap is 67%, which equates to non-overlap of 33%; that is, δ = .33. Note that since Cliff’s δ and Cohen’s d are on different scales, the original statistics are not directly comparable. However, both can be described in terms of the overlap of the distributions when normality is assumed, which allows conversion of Cliff’s δ into an unbiased estimate of the standardised mean difference via this bridge.
As noted earlier, there are relatively few statistical packages that calculate Cliff’s delta, or that offer other non-parametric effect size metrics. As a result, an Excel calculator was developed and tested as part of this thesis, to allow for calculation of the statistic. This program was used for many of the calculations of $\delta$ presented in this thesis, and is appended to the electronic version, but this calculator has practical limitations due to the computational constraints noted above, with considerable difficulties arising with larger sample sizes (greater than 500). Recently however, an R package called ‘orddom’ has been developed (Rogmann, 2013), which allows calculation of delta for any sample size – though users should note that sample sizes greater than 5000 require a computer with at least two gigabytes of unallocated RAM and several hours of calculation time. This package corroborated the estimates of delta provided by the Excel calculator, and allowed extension of the range of comparisons presented in the analysis section.

3.3. Analysis

The remainder of this chapter presents the results of analyses using several of the effect size procedures described above, and compares the results derived from each metric. The effect sizes selected for comparison include: standardised mean difference, mainly due to its popularity; odds ratios; and Cliff’s delta. The measure of SMD presented uses the Hedges’ bias correction as described above, since this provides the least biased estimate of effect size for parametric data.
The uncorrected equivalent (e.g., Cohen's \( d \)) is not presented for comparison since the degree of bias is exceedingly small for medium to large sample sizes, so would not provide additional information. Non-standardised group differences also provide little in the way of additional information since these can easily be derived, so this metric is not supplied.

As is the case throughout this thesis, all analyses are based on the 2006-2007 and 2008-2009 cohorts of the LPDP and, where possible, are based on identical groups of students in order to maximise comparability of results, both within and across chapters. However, as previously noted, this was not always possible due to the specific constraints of the particular analysis. In the current chapter, the analyses presented in the odds ratio section differ from those presented using SMD and \( \delta \) effect sizes, since ORs cannot be calculated for continuous data. As a result, the OR section is intended to be primarily illustrative and does not provide directly comparable results, but does provide an alternative insight into the effectiveness of the project for those students identified to be at greatest risk.

In order to standardise the scores across year levels, allowing for aggregation across year levels into the four separate cohorts of students, all analyses are based on asTTle scores after subtracting the expected score for each student's year level based on the asTTle normative sample. This facilitates interpretation of relative achievement levels by indicating when students are achieving above or below expectation, and also allows the effect sizes to be interpreted in terms of the value added by the project, since usual maturational progress has already been accounted for.

### 3.3.1. Analysis: Odds ratios

As noted in the introductory section to this chapter, odds ratios provide a useful effect size measure for binary categorical data. Since asTTle scores are effectively continuous, it is impossible to calculate an OR directly; to quantify the overall progress of students attending schools that participated in LPDP, for example. However, in many educational interventions, one might be interested to determine the comparative likelihood of one group meeting particular
Chapter 3 - Effect sizes

expectations compared with another and, in doing so, determine whether there are particular demographic groups where the intervention is more successful. This has been explored with the LPDP data by considering the progress among the lowest achieving students. This group is of particular interest for a number of reasons. Firstly, from an analytic point of view, this is the group for whom regression to the mean (discussed in Chapter 4) presents a particular problem for the interpretation of standardised group mean difference effect sizes. Secondly, since reductions in disparity among subgroups delineated by ethnicity and socio-economic status were particular aims of LPDP, determining the extent to which this has occurred among the group of students with the lowest baseline scores is an excellent way of quantifying how well this aim has been met, especially since there were large initial disparities in the make-up of this group.

The group of students with the lowest achievement has been defined as those whose baseline asTTle score was in the lowest 20% of students for their year level. Previous LPDP milestone reports have suggested that progress among these students was considerably greater than overall project progress; already noted to be generally higher than that of the wider New Zealand student population (as evidenced by the asTTle normative sample), but that a small portion of students with low baseline achievement fell further behind the expected norms for their age (Timperley et al., 2010). This section explores the underlying LPDP aim of disparity reduction by comparing the ORs of students in the lowest 20% to investigate the characteristics of students that were less well served by the project. It is difficult to ascertain precisely why some students have not made sufficient gains, but it is often useful to identify the characteristics of these students to investigate potential solutions. The gender, ethnicity, and decile group of students who have not made progress have been compared with other students in the lowest 20% who did make gains. Since ORs can only be calculated for binary categorical data, each comparison made reflects the identified target compared with all other students whose baseline scores were in the lowest 20%. For example, when comparing the odds of NZ European students, the comparison group includes students of all other ethnicities, including NZ Māori and Pasifika.

It should be noted that the majority of students whose baseline scores were in the lowest 20% of their year level did make gains significantly above normative expectations; approximately 7%
of students in the lowest 20% of the writing cohorts progressed slower than the normative expectations, while in reading schools, 25% of students with low baseline scores in the 2006–2007 cohort, and 12% of those in the 2008–2009 cohort gained less than expected for their year level. Table 2 shows the odds of each subgroup making larger gains than the normative sample. Conditional formatting is used to identify subgroups with comparatively better or worse odds, with green indicating that the odds of the particular subgroup making progress were good, and red indicating poor odds of progress in real terms. Note that in the majority of instances, the difference in odds was non-significant, so it is difficult to make conclusions about the relative efficacy of the project for these specific subgroups with any particular certainty. This is largely because odds ratios have limited power, and require fairly substantial sample sizes or large differences in odds in order to reliably separate the odds of two groups. In order to increase the sample size, and to determine whether particular subgroups had different overall odds of progress across the entire project, an overall column is included in Table 2, aggregating the odds of students with low baseline scores for the project overall. The overall odds of each subgroup of students in the lowest 20% making progress at or above expectation was relatively similar for most subgroups, with the exception of students of NZ Māori descent, who had lower odds of making progress.
Table 2. The odds of students with baseline scores in the lowest 20% progressing above asTTle expectations for their year group by subgroup.

<table>
<thead>
<tr>
<th>Subgroups</th>
<th>2006–07 cohort</th>
<th>2008–09 cohort</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reading</td>
<td>Writing</td>
<td>Reading</td>
</tr>
<tr>
<td>Māori</td>
<td>0.63**</td>
<td>0.62</td>
<td>0.94</td>
</tr>
<tr>
<td>Pasifika</td>
<td>1.16</td>
<td>0.67</td>
<td>1.15</td>
</tr>
<tr>
<td>NZ European</td>
<td>1.27</td>
<td>1.61</td>
<td>0.49*</td>
</tr>
<tr>
<td>Boys</td>
<td>0.91</td>
<td>0.63</td>
<td>0.47</td>
</tr>
<tr>
<td>Low decile</td>
<td>0.71*</td>
<td>1.02</td>
<td>0.95</td>
</tr>
<tr>
<td>Mid decile</td>
<td>1.01</td>
<td>0.54</td>
<td>0.85</td>
</tr>
<tr>
<td>High decile</td>
<td>1.77*</td>
<td>7.7*</td>
<td>1.45</td>
</tr>
</tbody>
</table>

* p < .05, ** p < .01, *** p < .001 (denotes significant difference in odds of progressing)

In both of the 2006–2007 cohorts (i.e., reading and writing), NZ Māori students in the lowest 20% had lower odds of meeting normative expectations of progress compared with other students with very low baseline scores, though this was non-significant in writing since relatively few students failed to progress in this domain. The odds of Māori students progressing at or above expectation improved in the 2008–2009 cohort and the odds were not markedly different from students of other ethnicities. Pasifika students in the lowest 20% appear to have slightly better odds of making progress than other students when their schools’ focus was reading, while the reverse was true in schools with a writing focus, though none of these odds was significantly different. Boys with low baseline scores were also generally less likely than girls to meet normative expectations of progress, though again this tendency was non-significant – even for boys in the 2008-2009 reading cohort in which the odds appear much lower (this cohort was smaller than the other three, and only 12% of students in the lowest 20% failed to progress).

Students in high-decile schools were comparatively unlikely to have baseline scores in the lowest 20% (overall, only 12% of students in the lowest 20% were attending high decile schools; compared with 43% each at low and mid decile schools), but the few that did have low initial achievement had significantly higher odds of making above expected progress during the 2006-
2007 cohort. The odds for high decile students in writing for the 2006-2007 cohort are substantially higher than for any other group because only one of the 95 students attending high decile schools with low baseline scores failed to progress at or above expectation. Conversely, students attending low decile schools had significantly lower odds of making progress in reading during the 2006-2007 cohort, and writing in the 2008-2009 cohort.

3.3.2. Analysis: Standardised Mean Difference versus Cliff’s delta

To help contextualise the data described, the first two tables in each section provide descriptive information about the specific cluster being described (i.e., cohort and focus), for each of the two project years. These tables include the number of students for whom longitudinal data were obtained during each year of the project, the mean and standard deviation of the scores for each subgroup at each measurement point, along with information about the degree of departure from normality. As discussed in Chapter 2, achievement data were collected at three measurement points; the beginning of the first year in the project (baseline), the end of the first year (EOY1) and the end of the second year (EOY2).

The degree of non-normality is assessed using the Kolmogorov-Smirnov test (with the Lilliefors Significance Correction since no assumptions are made about the sample means), which quantifies the extent to which the data being tested depart from the normal distribution, denoted as D (Razali & Wah, 2011). A D statistic of 0 indicates that the data demonstrate perfect normality while larger values represent increasingly large departures from normality. Significance levels for this test are dependent on sample size, meaning that in large samples even small departures from normality result in statistical significance being attributed to the departure (Field & Miles, 2010), so the absolute value is often more informative. Indeed, there are authors that argue that normality tests are of limited utility since when sample size is small, power to detect non-normality is low, and when sample size is large, even minor departures from normality are labelled significant (D’Agostino, 1986). However, in this instance, the intention is simply to provide an indication of the relative degree of non-normality present in the data for contextualisation.
purposes. To demonstrate visually the degree of departure associated with the LPDP data, a histogram is shown below, representing the most extreme overall departure from normality (D = .068, p < .001). The histogram in Figure 3 indicates some kurtosis around the centre of the distribution, and skew toward lower extreme scores.

![Figure 3. Histogram of T1 Score against expectation for Cohort 2 Writing.](image)

Figure 4 represents the departure from normality in the above example as a Q-Q plot, showing more clearly the departure from normality at both extremes.
To assist with interpretation, these contextualisation tables are coloured using conditional formatting so that results can be compared holistically without needing to compare each of the numbers in the tables individually. Colours range from red to yellow to green, with red indicating a poor result, yellow an average result and green, a positive one. To maximise comparability, the colours used in the contextualisation tables have the same meaning for the four different cohorts (or clusters) of students. For the mean scores, which indicate the average difference between the scores of the specific group and those of similar students in the normative sample (i.e., of the same age and schooling system), deep red indicates achievement equal to or more than half a standard deviation unit (SDU) below the normative sample. Deep green indicates achievement half a standard deviation higher than students in the normative sample and yellow indicates achievement approximately equal to expectation. With respect to the results of the Kolmogorov Smirnov test (K-S D), green indicates no departure from normality, yellow a modest departure (K-S D = .05) and red, notable departure (K-S D = .1). However, it should be noted that Micceri (1989) found that an extensive review of 440 psychometric and achievement distributions with moderate to large sample sizes (70% of these had sample sizes larger than 1000) showed that all of these distributions were significantly non-normal according to the Kolmogorov Smirnov test. In addition, all of the distributions analysed throughout this thesis showed modest skewness (< 1 in
all cases) and kurtosis (< 2 in all cases). Though definitions differ, according to Kline (2005), skewness levels less than three, or kurtosis less than ten, suggest departures from normality that are unlikely to be problematic, while West, Finch and Curran (1995) and Kim (2013) suggest slightly more conservative criteria, deeming skewness less than two and kurtosis less than seven to be acceptable.

The third table in each of the following sections compares the relative magnitude of effect sizes calculated using SMD against those calculated using $\delta$. Since the reader is more likely to be familiar with the magnitude of SMD effect sizes, $\delta$ is also converted into the same scale as SMD, using the degree of distributional overlap, hereafter referred to as the non-parametric standardised mean difference, or NPSMD. This is especially useful since, as noted previously, the SMD and the NPSMD should be equivalent when the data are normally distributed. Any discrepancies show the implications of ignoring the normality assumptions inherent in the SMD calculation. In order to make the differences clearer, the tables are again coloured using conditional formatting, with deep red indicating the lowest effect size, yellow the average effect, and deep green the largest effect for the cluster. The progress of each cluster differs markedly for various reasons, most notably baseline achievement, so in the progress tables the colours are contextualised within each cluster. In addition, since the effect sizes for the overall project (i.e., a two-year period) should be larger than the one-year effects, this is taken into account. This means that effect sizes marked as deep green for reading schools in cohort 1 are smaller than those marked green in other clusters, but enhances comparison of the SMD with the NPSMD. Since the focus of this section centres on this comparison, and the scale for $\delta$ differs from that of the SMD and NPSMD, no colour is used for the $\delta$ column.

Note that all effect sizes presented reflect the “added value” of the project – since achievement is measured with respect to the distance from the normative sample achievement, normal maturational progress has already been taken into account.
3.3.2.1. Reading Focus – 2006-2007 Cohort

Table 3 presents descriptive statistics for the first year of the 2006-2007 cohort. Overall, this cohort had average baseline scores somewhat higher than those in the asTTle normative sample – the difference of almost 20 points equates to an effect size approximately equal to 0.2 since the standard deviation of the tool is assumed to be 100. Average scores at baseline follow the typical pattern of results by demographic, with boys, students attending schools with a lower socio-economic catchment area, and minority groups exhibiting lower average scores. There appears to be a small decline in the average scores relative to those in the normative sample by the end of the first year of the project; subgroups typically are represented with a less deep green, and in some cases darker red colour by EOY1. In almost all cases, the Kolmogorov-Smirnov tests (with Lilliefors Correction) show that the data do not meet normality assumptions, though in absolute terms these departures would typically be considered modest.
Table 3. Descriptive statistics for the first year of the project: Reading 2006-2007.

<table>
<thead>
<tr>
<th>Factor</th>
<th>N</th>
<th>Baseline mean (SD)</th>
<th>EOY1 mean (SD)</th>
<th>K-S D Baseline</th>
<th>K-S D EOY1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>5196</td>
<td>29.4 (79.9)</td>
<td>19.4 (82.5)</td>
<td>.032***</td>
<td>.030***</td>
</tr>
<tr>
<td>Male</td>
<td>2728</td>
<td>20.7 (81.2)</td>
<td>10.4 (83.6)</td>
<td>.037***</td>
<td>.039***</td>
</tr>
<tr>
<td>Female</td>
<td>2455</td>
<td>39.1 (77.5)</td>
<td>29.3 (80.1)</td>
<td>.028***</td>
<td>.029***</td>
</tr>
<tr>
<td>Low Decile</td>
<td>1124</td>
<td>4.1 (76.9)</td>
<td>-3.5 (80.3)</td>
<td>.050***</td>
<td>.039**</td>
</tr>
<tr>
<td>Mid Decile</td>
<td>2211</td>
<td>18.5 (78.5)</td>
<td>7.9 (82.7)</td>
<td>.038***</td>
<td>.033***</td>
</tr>
<tr>
<td>High Decile</td>
<td>1848</td>
<td>58.4 (74.4)</td>
<td>47.5 (75.7)</td>
<td>.032***</td>
<td>.038***</td>
</tr>
<tr>
<td>NZ Euro</td>
<td>3262</td>
<td>42.9 (79.9)</td>
<td>34.5 (79.4)</td>
<td>.033***</td>
<td>.032***</td>
</tr>
<tr>
<td>NZ Māori</td>
<td>1016</td>
<td>-6.8 (73.7)</td>
<td>-19.7 (76.6)</td>
<td>.035**</td>
<td>.030*</td>
</tr>
<tr>
<td>Pasifika</td>
<td>270</td>
<td>-21.4 (73.5)</td>
<td>-36.3 (70.8)</td>
<td>.070**</td>
<td>.033</td>
</tr>
<tr>
<td>Other</td>
<td>635</td>
<td>39.2 (81.5)</td>
<td>25.6 (81.5)</td>
<td>.060***</td>
<td>.052***</td>
</tr>
</tbody>
</table>

Table 4 presents descriptive statistics for the second year of the 2006-2007 cohort. The very slight differences between the EOY1 means in this table and those presented for the first year of the project indicate that the EOY1 scores were very slightly higher (~2-4 points) among students retained into the second year of the project. There is a clear shift toward higher achievement during the second year of the project among all subgroups, with most groups at least half a standard deviation above the mean at EOY2 (those coloured deep green). Indeed, all subgroups are achieving at or above expectation by the EOY2 assessment - though a considerable gap remains between students of NZ Māori and Pasifika descent and those of NZ European or other ethnicities. The results of the Kolmogorov-Smirnov test (with Lilliefors Correction) suggest moderate non-normality among certain subgroups at the EOY2 assessment.
Table 4. Descriptive statistics for the second year of the project: Reading 2006-2007.

<table>
<thead>
<tr>
<th>Factor</th>
<th>N</th>
<th>EOY1 mean (SD)</th>
<th>EOY2 mean (SD)</th>
<th>K-S D EOY1</th>
<th>K-S D EOY2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>3881</td>
<td>21.7 (81.9)</td>
<td>54.5 (97.4)</td>
<td>.030***</td>
<td>.034***</td>
</tr>
<tr>
<td>Male</td>
<td>2027</td>
<td>12.1 (83.0)</td>
<td>46.2 (97.4)</td>
<td>.041***</td>
<td>.024**</td>
</tr>
<tr>
<td>Female</td>
<td>1834</td>
<td>32.7 (97.4)</td>
<td>64.6 (95.9)</td>
<td>.026**</td>
<td>.052***</td>
</tr>
<tr>
<td>Low Decile</td>
<td>873</td>
<td>0.1 (82.6)</td>
<td>23.9 (102.6)</td>
<td>.044***</td>
<td>.070***</td>
</tr>
<tr>
<td>Mid Decile</td>
<td>1708</td>
<td>10.9 (80.7)</td>
<td>51.8 (97.4)</td>
<td>.031***</td>
<td>.032**</td>
</tr>
<tr>
<td>High Decile</td>
<td>1280</td>
<td>51.5 (73.7)</td>
<td>80.3 (85.5)</td>
<td>.039***</td>
<td>.060***</td>
</tr>
<tr>
<td>NZ Euro</td>
<td>2442</td>
<td>37.9 (77.8)</td>
<td>71.1 (96.1)</td>
<td>.028***</td>
<td>.033***</td>
</tr>
<tr>
<td>NZ Māori</td>
<td>761</td>
<td>-18.4 (78.6)</td>
<td>7.5 (94.1)</td>
<td>.039**</td>
<td>.054***</td>
</tr>
<tr>
<td>Pasifika</td>
<td>202</td>
<td>-37.7 (65.9)</td>
<td>9.0 (79.4)</td>
<td>.042</td>
<td>.059</td>
</tr>
<tr>
<td>Other</td>
<td>456</td>
<td>29.6 (78.5)</td>
<td>67.7 (81.3)</td>
<td>.055**</td>
<td>.060***</td>
</tr>
</tbody>
</table>

Legend

<table>
<thead>
<tr>
<th>Achievement</th>
<th>Colour</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.5SDU</td>
<td>Red</td>
</tr>
<tr>
<td>At norm</td>
<td>Green</td>
</tr>
<tr>
<td>+0.5SDU</td>
<td>Yellow</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Departure from normality</th>
</tr>
</thead>
<tbody>
<tr>
<td>D = .1</td>
</tr>
<tr>
<td>D = .05</td>
</tr>
<tr>
<td>D = 0</td>
</tr>
</tbody>
</table>

*p < .05, ** p < .01, *** p < .001 (denotes significant departure from normality)

Progress during each project year for schools in the reading focus of the 2006-2007 cohort is shown in Table 5. Progress was comparatively low among all subgroups during the first year of the project, with the absolute values indicating slower progress than that of the normative sample, while progress in the second year was considerably more accelerated, especially among Pasifika. Overall progress was lower among students in low decile schools and those of NZ Māori descent. Conclusions about which subgroups progressed the most rapidly were the same for the NPSMD as for the SMD, but the SMD effect size estimates showed considerable bias, with an inflation factor ranging from 10 to 220%, with an average of 50%. This means that, in some instances, a considerable portion of the apparent effect measured using SMD was as a result of the distributional properties rather than a real effect. For example, the NPSMD indicates an overall effect for students in low decile schools 0.05 larger than for the normative sample – an effect of
little practical significance. Conversely, the SMD effect size is 0.16 – not a large effect, but large enough that some gain might be claimed.

### Table 5. Comparison of NPSMD and SMD estimates of progress effect sizes in each year of the project, and for the project overall: Reading 2006-2007.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Year 1 Progress</th>
<th>Year 2 Progress</th>
<th>Overall Progress</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NPSMD</td>
<td>SMD</td>
<td>NPSMD</td>
</tr>
<tr>
<td>Overall</td>
<td>-0.07</td>
<td>-0.08</td>
<td>-0.12</td>
</tr>
<tr>
<td>Male</td>
<td>-0.07</td>
<td>-0.08</td>
<td>-0.12</td>
</tr>
<tr>
<td>Female</td>
<td>-0.07</td>
<td>-0.08</td>
<td>-0.12</td>
</tr>
<tr>
<td>Low Decile</td>
<td>-0.06</td>
<td>-0.08</td>
<td>-0.09</td>
</tr>
<tr>
<td>Mid Decile</td>
<td>-0.07</td>
<td>-0.09</td>
<td>-0.13</td>
</tr>
<tr>
<td>High Decile</td>
<td>-0.08</td>
<td>-0.09</td>
<td>-0.13</td>
</tr>
<tr>
<td>NZ Euro</td>
<td>-0.05</td>
<td>-0.06</td>
<td>-0.1</td>
</tr>
<tr>
<td>NZ Māori</td>
<td>-0.09</td>
<td>-0.11</td>
<td>-0.16</td>
</tr>
<tr>
<td>Pasifika</td>
<td>-0.14</td>
<td>-0.16</td>
<td>-0.2</td>
</tr>
<tr>
<td>Other</td>
<td>-0.1</td>
<td>-0.12</td>
<td>-0.16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Colour Legend</th>
<th>Progress relative to other subgroups in cohort¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lowest</td>
<td>Average</td>
</tr>
</tbody>
</table>

¹ Note that the colours for overall effects (two years) are calibrated separately from single year effects.

### 3.3.2.2. Reading Focus – 2008-2009 Cohort

The baseline scores of students attending schools with a reading focus of the 2008-2009 cohort showed considerably more variability by subgroup than the previous reading cohort, as shown in Table 6 below, but the pattern was similar, with students in typically at risk groups more likely to have baseline scores lower than normative expectation. Students attending high decile schools, and those of NZ European descent, had baseline scores that were already considerably
higher than expectation. Subgroups with lower average achievement typically had higher levels of non-normality – unsurprising since such groups typically show greater skew, with an increased likelihood of very low scores affecting the overall subgroup mean.

Table 6. Descriptive statistics for the first year of the project: Reading 2008-2009.

<table>
<thead>
<tr>
<th>Factor</th>
<th>N</th>
<th>Baseline mean (SD)</th>
<th>EOY1 mean (SD)</th>
<th>K-S D Baseline</th>
<th>K-S D EOY1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>2444</td>
<td>12.2 (98.0)</td>
<td>25.0 (88.0)</td>
<td>.039***</td>
<td>.039***</td>
</tr>
<tr>
<td>Male</td>
<td>1201</td>
<td>-0.1 (99.2)</td>
<td>12.0 (88.7)</td>
<td>.033**</td>
<td>.044***</td>
</tr>
<tr>
<td>Female</td>
<td>1243</td>
<td>24.1 (88.0)</td>
<td>37.6 (85.4)</td>
<td>.046***</td>
<td>.044***</td>
</tr>
<tr>
<td>Low Decile</td>
<td>1015</td>
<td>-37.0 (86.4)</td>
<td>-1.3 (86.4)</td>
<td>.077***</td>
<td>.058***</td>
</tr>
<tr>
<td>Mid Decile</td>
<td>516</td>
<td>5.7 (76.6)</td>
<td>10.5 (76.5)</td>
<td>.060***</td>
<td>.054**</td>
</tr>
<tr>
<td>High Decile</td>
<td>913</td>
<td>70.5 (89.6)</td>
<td>62.5 (82.6)</td>
<td>.031*</td>
<td>.069***</td>
</tr>
<tr>
<td>NZ Euro</td>
<td>1107</td>
<td>48.0 (94.7)</td>
<td>47.8 (87.7)</td>
<td>.036**</td>
<td>.057***</td>
</tr>
<tr>
<td>NZ Māori</td>
<td>608</td>
<td>-29.8 (83.9)</td>
<td>-2.4 (77.9)</td>
<td>.079***</td>
<td>.053***</td>
</tr>
<tr>
<td>Pasifika</td>
<td>335</td>
<td>-32.2 (80.8)</td>
<td>-12.1 (80.1)</td>
<td>.067***</td>
<td>.073***</td>
</tr>
<tr>
<td>Other</td>
<td>394</td>
<td>14.1 (100.3)</td>
<td>34.8 (87.8)</td>
<td>.046*</td>
<td>.050*</td>
</tr>
</tbody>
</table>

* p < .05, ** p < .01, *** p < .001 (denotes significant departure from normality)

Table 7 indicates that by the end of the second year of the project, all groups in the 2008-2009 reading cohort were achieving at or above normative expectation. However, considerable variability remained, with an average score difference equal to three-quarters of a standard deviation between students in low decile and those in high decile schools. There was a notable difference in the average scores for the EOY1 assessment among students tracked longitudinally during the second year, with the average scores for these students approximately 20 points lower at EOY1 (see Table 7) than students tracked longitudinally during the first year (see Table 6). Small differences should be expected, since the schooling structure means that not all students
were able to be retained into the second year of the project, but the differences in this case warrant further investigation. It was determined that the differences were, indeed, systemic, and related to the loss of Year 8 students at the end of the first year. Students in Year 8 had especially high rates of progress during the first year of this cohort, lifting overall EOY1 scores for students tracked over year one of the project. The greatest levels of non-normality at the EOY2 assessment were among relatively high achievement groups. It appears that these groups had greater rates of progress in the second year of the project (see Table 7), but that this generated some skewness in the tail.

Table 7. Descriptive statistics for the second year of the project: Reading 2008-2009.

<table>
<thead>
<tr>
<th>Factor</th>
<th>N</th>
<th>EOY1 mean (SD)</th>
<th>EOY2 mean (SD)</th>
<th>K-S D EOY1</th>
<th>K-S D EOY2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>1547</td>
<td>7.7 (80.3)</td>
<td>35.0 (96.7)</td>
<td>.036***</td>
<td>.055***</td>
</tr>
<tr>
<td>Male</td>
<td>754</td>
<td>-3.3 (81.9)</td>
<td>25.4 (94.6)</td>
<td>.043**</td>
<td>.055***</td>
</tr>
<tr>
<td>Female</td>
<td>753</td>
<td>19.3 (96.7)</td>
<td>45.8 (97.0)</td>
<td>.03</td>
<td>.062***</td>
</tr>
<tr>
<td>Low Decile</td>
<td>690</td>
<td>-13.1 (81.9)</td>
<td>4.7 (93.0)</td>
<td>.052***</td>
<td>.058***</td>
</tr>
<tr>
<td>Mid Decile</td>
<td>317</td>
<td>15.3 (75.3)</td>
<td>32.1 (73.6)</td>
<td>.076***</td>
<td>0.039</td>
</tr>
<tr>
<td>High Decile</td>
<td>500</td>
<td>32.4 (71.8)</td>
<td>80.5 (96.3)</td>
<td>.053**</td>
<td>0.081***</td>
</tr>
<tr>
<td>NZ Euro</td>
<td>650</td>
<td>27.9 (78.0)</td>
<td>58.3 (91.0)</td>
<td>.057***</td>
<td>0.056***</td>
</tr>
<tr>
<td>NZ Māori</td>
<td>329</td>
<td>-10.6 (70.6)</td>
<td>4.1 (84.7)</td>
<td>0.036</td>
<td>0.048</td>
</tr>
<tr>
<td>Pasifika</td>
<td>277</td>
<td>-21.4 (75.7)</td>
<td>-1.9 (90.3)</td>
<td>0.070**</td>
<td>0.052</td>
</tr>
<tr>
<td>Other</td>
<td>251</td>
<td>13.1 (85.5)</td>
<td>59.7 (105.8)</td>
<td>0.056</td>
<td>0.087***</td>
</tr>
</tbody>
</table>

Colour | Achievement | Departure from normality
--------|-------------|------------------------
| -0.5SDU | At norm | +0.5SDU
Legend

* p < .05, ** p < .01, *** p < .001 (denotes significant departure from normality)

Progress in the first year of the project was generally modest, accelerating in the second year. Overall gains were relatively consistent across subgroups, with higher gains among students attending high decile schools, and those of “other” ethnicities. There is again considerable
discrepancy between the NPSMD and the SMD, with an average inflation factor of 55% for the SMD estimate of effect size, ranging from 25-100%. In absolute terms, these differences are in some cases large enough that different conclusions would be made with respect to practical significance of the gains. For example, the overall SMD for Pasifika is 0.38, which is almost double the unbiased NPSMD estimate of the effect of 0.2. In addition, the SMD effect for Pasifika is the same as for NZ European students, which would lead researchers to conclude that progress was equivalent for both groups, when distributional shift indicates that the shift among Pasifika was notably lower.

Table 8. Comparison of NPSMD and SMD estimates of progress effect sizes in each year of the project, and for the project overall: Reading 2008-2009.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Year 1 Progress</th>
<th>Year 2 Progress</th>
<th>Overall Progress</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>δ</td>
<td>NPSMD</td>
<td>SMD</td>
</tr>
<tr>
<td>Overall</td>
<td>0.06</td>
<td>0.08</td>
<td>0.13</td>
</tr>
<tr>
<td>Male</td>
<td>0.06</td>
<td>0.08</td>
<td>0.12</td>
</tr>
<tr>
<td>Female</td>
<td>0.07</td>
<td>0.09</td>
<td>0.14</td>
</tr>
<tr>
<td>Low Decile</td>
<td>0.18</td>
<td>0.26</td>
<td>0.41</td>
</tr>
<tr>
<td>Mid Decile</td>
<td>0.03</td>
<td>0.06</td>
<td>0.11</td>
</tr>
<tr>
<td>High Decile</td>
<td>-0.06</td>
<td>-0.07</td>
<td>-0.09</td>
</tr>
<tr>
<td>NZ Euro</td>
<td>-0.01</td>
<td>-0.01</td>
<td>0</td>
</tr>
<tr>
<td>NZ Māori</td>
<td>0.14</td>
<td>0.19</td>
<td>0.33</td>
</tr>
<tr>
<td>Pasifika</td>
<td>0.1</td>
<td>0.13</td>
<td>0.24</td>
</tr>
<tr>
<td>Other</td>
<td>0.11</td>
<td>0.15</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Colour Legend: Progress relative to other subgroups in cohort¹

¹ Note that the colours for overall effects (two years) are calibrated separately from single year effects.

3.3.2.3. Writing Focus – 2006-2007 Cohort

Students attending schools with a writing focus during the 2006-2007 cohort had baseline scores considerably lower than students in the normative sample, with the exception of high
decile students, whose scores were similar to expectation. Considerable shift is evident across the first year as shown in Table 9. The degree of non-normality is fairly consistent over time and among subgroups, with the exception of the baseline scores for NZ Māori.


<table>
<thead>
<tr>
<th>Factor</th>
<th>N</th>
<th>Baseline mean (SD)</th>
<th>EOY1 mean (SD)</th>
<th>K-S D Baseline</th>
<th>K-S D EOY2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>5051</td>
<td>-44.5 (123.6)</td>
<td>-7.3 (112.4)</td>
<td>.056***</td>
<td>.056***</td>
</tr>
<tr>
<td>Male</td>
<td>2569</td>
<td>-70.5 (124.6)</td>
<td>-26.7 (114.2)</td>
<td>.066***</td>
<td>.060***</td>
</tr>
<tr>
<td>Female</td>
<td>2475</td>
<td>-17.8 (112.4)</td>
<td>13.0 (106.8)</td>
<td>.046***</td>
<td>.060***</td>
</tr>
<tr>
<td>Low Decile</td>
<td>2229</td>
<td>-78.3 (120.2)</td>
<td>-26.0 (113.1)</td>
<td>.066***</td>
<td>.054***</td>
</tr>
<tr>
<td>Mid Decile</td>
<td>1464</td>
<td>-34.3 (125.5)</td>
<td>-18.6 (113.0)</td>
<td>.061***</td>
<td>.054***</td>
</tr>
<tr>
<td>High Decile</td>
<td>1351</td>
<td>-0.3 (110.5)</td>
<td>35.9 (98.3)</td>
<td>.062***</td>
<td>.059***</td>
</tr>
<tr>
<td>NZ Euro</td>
<td>2713</td>
<td>-31.9 (122.8)</td>
<td>1.0 (109.5)</td>
<td>.055***</td>
<td>.056***</td>
</tr>
<tr>
<td>NZ Māori</td>
<td>1176</td>
<td>-70.8 (120.8)</td>
<td>-31.1 (120.1)</td>
<td>.082***</td>
<td>.060***</td>
</tr>
<tr>
<td>Pasifika</td>
<td>548</td>
<td>-62.7 (121.5)</td>
<td>-8.6 (108.0)</td>
<td>.067***</td>
<td>.053***</td>
</tr>
<tr>
<td>Other</td>
<td>607</td>
<td>-34.8 (124.5)</td>
<td>3.3 (106.9)</td>
<td>.043**</td>
<td>.063***</td>
</tr>
</tbody>
</table>

* p < .05, ** p < .01, *** p < .001 (denotes significant departure from normality)

By the end of the second year of the cohort, all subgroups had average scores similar to, or greater than, the normative sample (see Table 10). Considerable differences in outcome remain by subgroup however, especially with respect to school decile level. The relatively lower levels of non-normality at the EOY2 measurement reflect a reduction in the left skew; that is, there were fewer students with extremely low scores. There is no apparent reduction in variance due to the floor effect in the asTTle writing tool evident in this cohort of students.

<table>
<thead>
<tr>
<th>Factor</th>
<th>N</th>
<th>EOY1 mean (SD)</th>
<th>EOY2 mean (SD)</th>
<th>K-S D EOY1</th>
<th>K-S D EOY2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>3550</td>
<td>-12.1 (113.7)</td>
<td>25.2 (110.4)</td>
<td>.058***</td>
<td>.048***</td>
</tr>
<tr>
<td>Male</td>
<td>1780</td>
<td>-31.4 (116.2)</td>
<td>4.4 (108.5)</td>
<td>.061***</td>
<td>.057***</td>
</tr>
<tr>
<td>Female</td>
<td>1748</td>
<td>7.6 (110.4)</td>
<td>45.9 (108.3)</td>
<td>.055***</td>
<td>.041***</td>
</tr>
<tr>
<td>Low Decile</td>
<td>1490</td>
<td>-35.8 (115.8)</td>
<td>-5.9 (106.6)</td>
<td>.056***</td>
<td>.064***</td>
</tr>
<tr>
<td>Mid Decile</td>
<td>1031</td>
<td>-24.4 (112.8)</td>
<td>19.4 (118.4)</td>
<td>.059***</td>
<td>.035**</td>
</tr>
<tr>
<td>High Decile</td>
<td>1007</td>
<td>35.7 (96.1)</td>
<td>76.3 (86.3)</td>
<td>.045***</td>
<td>.029*</td>
</tr>
<tr>
<td>NZ Euro</td>
<td>1900</td>
<td>-3.5 (110.2)</td>
<td>37.4 (107.8)</td>
<td>.052***</td>
<td>.044***</td>
</tr>
<tr>
<td>NZ Māori</td>
<td>833</td>
<td>-41.8 (121.3)</td>
<td>-10.4 (117.6)</td>
<td>.067***</td>
<td>.061***</td>
</tr>
<tr>
<td>Pasifika</td>
<td>349</td>
<td>-15.2 (113.4)</td>
<td>7.1 (100.6)</td>
<td>.075***</td>
<td>0.043</td>
</tr>
<tr>
<td>Other</td>
<td>446</td>
<td>9.5 (103.7)</td>
<td>52.0 (95.5)</td>
<td>.057**</td>
<td>.049*</td>
</tr>
</tbody>
</table>

* p < .05, ** P < .01, *** p < .001 (denotes significant departure from normality)

In absolute terms, the effect size estimates were much larger for students in this cohort than for either of the reading cohorts. However, the colour pattern demonstrates that there was considerable variability in the size of these shifts among different subgroups, as well as between the NPSMD and SMD estimates (see Table 11). Overall, progress was lower among students attending mid-decile schools, and marginally higher among Pasifika students. Once again, the discrepancy between the NPSMD and the SMD is immediately visible, though the inflation factor for the SMD estimate of effect size is much smaller, ranging from 15-55%, with an average inflation of 25%. However, the degree of inflation present in this cohort is generally unlikely to have major implications for conclusions – although all of the SMD estimates are larger than the unbiased estimates, the absolute difference is typically modest.
Table 11. Comparison of NPSMD and SMD estimates of progress effect sizes in each year of the project, and for the project overall: Writing 2006-2007.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Year 1 Progress</th>
<th>Year 2 Progress</th>
<th>Overall Progress</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>δ</td>
<td>NPSMD</td>
<td>SMD</td>
</tr>
<tr>
<td>Overall</td>
<td>0.17</td>
<td>0.24</td>
<td>0.31</td>
</tr>
<tr>
<td>Male</td>
<td>0.2</td>
<td>0.28</td>
<td>0.36</td>
</tr>
<tr>
<td>Female</td>
<td>0.15</td>
<td>0.21</td>
<td>0.27</td>
</tr>
<tr>
<td>Low Decile</td>
<td>0.24</td>
<td>0.35</td>
<td>0.44</td>
</tr>
<tr>
<td>Mid Decile</td>
<td>0.07</td>
<td>0.09</td>
<td>0.13</td>
</tr>
<tr>
<td>High Decile</td>
<td>0.18</td>
<td>0.26</td>
<td>0.34</td>
</tr>
<tr>
<td>NZ Euro</td>
<td>0.15</td>
<td>0.21</td>
<td>0.28</td>
</tr>
<tr>
<td>NZ Māori</td>
<td>0.18</td>
<td>0.26</td>
<td>0.32</td>
</tr>
<tr>
<td>Pasifika</td>
<td>0.25</td>
<td>0.37</td>
<td>0.47</td>
</tr>
<tr>
<td>Other</td>
<td>0.17</td>
<td>0.23</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Colour Legend: Progress relative to other subgroups in cohort

1 Note that the colours for overall effects (two years) are calibrated separately from single year effects.

3.3.2.4. Writing Focus – 2008-2009 Cohort

The baseline scores of students whose schools chose to focus on writing in the 2008-2009 cohort of LPDP were much lower than students in the normative sample among all subgroups, with the baseline average scores of many subgroups close to, or greater than, a full standard deviation lower than students in the asTTle normative sample (see Table 12). By the end of the first year of the project, several groups were near expectation, while all groups show considerable improvement. Note that since the red in the table is used when average scores are half a standard deviation below expectation, the colour used for several subgroups (i.e., males, low decile, NZ Māori and Pasifika) did not change much despite a catch-up with respect to the normative sample of around half a standard deviation during the first year. The degree of non-
normality is relatively high at baseline, partly due to the floor effect associated with the asTTle writing tool.

Table 12. *Descriptive statistics for the first year of the project: Writing 2008-2009.*

<table>
<thead>
<tr>
<th>Factor</th>
<th>N</th>
<th>Baseline mean (SD)</th>
<th>EOY1 mean (SD)</th>
<th>K-S D Baseline</th>
<th>K-S D EOY1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>5433</td>
<td>-84.0 (128.3)</td>
<td>-25.5 (115.8)</td>
<td>.068***</td>
<td>.053***</td>
</tr>
<tr>
<td>Male</td>
<td>2770</td>
<td>-109.3 (128.6)</td>
<td>-48.0 (117.5)</td>
<td>.075***</td>
<td>.051***</td>
</tr>
<tr>
<td>Female</td>
<td>2653</td>
<td>-57.7 (115.8)</td>
<td>-1.9 (109.3)</td>
<td>.058***</td>
<td>.059***</td>
</tr>
<tr>
<td>Low Decile</td>
<td>1772</td>
<td>-122.3 (126.6)</td>
<td>-66.9 (125.5)</td>
<td>.079***</td>
<td>.060***</td>
</tr>
<tr>
<td>Mid Decile</td>
<td>2285</td>
<td>-82.5 (130.1)</td>
<td>-15.2 (106.5)</td>
<td>.065***</td>
<td>.050***</td>
</tr>
<tr>
<td>High Decile</td>
<td>1366</td>
<td>-37.0 (110.6)</td>
<td>11.1 (100.8)</td>
<td>.049***</td>
<td>.025*</td>
</tr>
<tr>
<td>NZ Euro</td>
<td>2630</td>
<td>-61.4 (123.4)</td>
<td>-6.0 (106.4)</td>
<td>.063***</td>
<td>.044***</td>
</tr>
<tr>
<td>NZ Māori</td>
<td>1031</td>
<td>-106.1 (133.8)</td>
<td>-42.8 (122.5)</td>
<td>.076***</td>
<td>.055***</td>
</tr>
<tr>
<td>Pasifika</td>
<td>863</td>
<td>-128.3 (117.3)</td>
<td>-69.5 (116.8)</td>
<td>.095***</td>
<td>.070***</td>
</tr>
<tr>
<td>Other</td>
<td>899</td>
<td>-82.7 (131.7)</td>
<td>-20.2 (119.5)</td>
<td>.079***</td>
<td>.058***</td>
</tr>
</tbody>
</table>

* p < .05, ** p < .01, *** p < .001 (denotes significant departure from normality)

There was a notable difference in the average scores for the EOY1 assessment among students tracked longitudinally during the second year, with the average scores for these students approximately 20 points lower at EOY1 (see Table 13) than students tracked longitudinally during the first year (see Table 12). As with the 2008-2009 reading cohort, these differences were related to the loss of Year 8 students at the end of the first year. Students in Years 7 and 8 had the highest average EOY1 scores (i.e., intermediate students), so the loss of the Year 8 students at the end of the first year of the project meant that those retained into the second year had apparently lower initial scores. By the end of the second year of the project, the overall cohort was achieving at expectation, while all groups show considerable improvement.
However, considerable discrepancies remain, with boys, students in low decile schools and those of NZ Māori and Pasifika descent continuing to achieve below expectation. The degree of non-normality is moderate for both measurement points.


<table>
<thead>
<tr>
<th>Factor</th>
<th>N</th>
<th>EOY1 mean (SD)</th>
<th>EOY2 mean (SD)</th>
<th>K-S D EOY1</th>
<th>K-S D EOY2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>3645</td>
<td>-45.1 (117.2)</td>
<td>0.0 (112.0)</td>
<td>.062***</td>
<td>.063***</td>
</tr>
<tr>
<td>Male</td>
<td>1766</td>
<td>-62.6 (116.7)</td>
<td>-18.2 (113.5)</td>
<td>.061***</td>
<td>.054***</td>
</tr>
<tr>
<td>Female</td>
<td>1730</td>
<td>-26.8 (112.0)</td>
<td>19.6 (105.8)</td>
<td>.072***</td>
<td>.066***</td>
</tr>
<tr>
<td>Low Decile</td>
<td>1132</td>
<td>-84.8 (127.5)</td>
<td>-32.9 (118.2)</td>
<td>.074***</td>
<td>.062***</td>
</tr>
<tr>
<td>Mid Decile</td>
<td>1438</td>
<td>-33.0 (105.4)</td>
<td>11.2 (109.9)</td>
<td>.049***</td>
<td>.059***</td>
</tr>
<tr>
<td>High Decile</td>
<td>926</td>
<td>-14.4 (105.5)</td>
<td>24.6 (94.1)</td>
<td>.050***</td>
<td>.051***</td>
</tr>
<tr>
<td>NZ Euro</td>
<td>1717</td>
<td>-25.5 (103.7)</td>
<td>21.8 (103.9)</td>
<td>.047***</td>
<td>.057***</td>
</tr>
<tr>
<td>NZ Māori</td>
<td>696</td>
<td>-65.2 (125.9)</td>
<td>-21.9 (117.3)</td>
<td>.062***</td>
<td>.060***</td>
</tr>
<tr>
<td>Pasifika</td>
<td>556</td>
<td>-80.7 (121.8)</td>
<td>-40.0 (110.5)</td>
<td>.086***</td>
<td>.080***</td>
</tr>
<tr>
<td>Other</td>
<td>527</td>
<td>-43.1 (124.5)</td>
<td>3.4 (110.3)</td>
<td>.063***</td>
<td>.072***</td>
</tr>
</tbody>
</table>

* p < .05, ** p < .01, *** p < .001 (denotes significant departure from normality)

The effect size estimates for this cohort were the largest of the four clusters. The colour pattern indicates considerable variation in the magnitude of these shifts among different subgroups, as well as between the NPSMD and SMD estimates (see Table 14). Overall, progress was lower among students attending high decile schools, and marginally lower among NZ Māori and Pasifika students. Yet again, the discrepancy between the NPSMD and the SMD is immediately visible, but the inflation factor for the SMD estimate of effect size is even smaller, ranging from 5-35%, with an average inflation of 20%. In this cohort, although there is a clear tendency for SMD estimates to be larger, these discrepancies are unlikely to have major
implications for conclusions; since the effect sizes are already relatively large, the conclusion
does not change, despite the inflation associated with the SMD.

Table 14. *Comparison of NPSMD and SMD estimates of progress effect sizes in each year of the project, and for the project overall: Writing 2008-2009.*

<table>
<thead>
<tr>
<th>Factor</th>
<th>Year 1 Progress</th>
<th>Year 2 Progress</th>
<th>Overall Progress</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>δ NPSMD SMD</td>
<td>δ NPSMD SMD</td>
<td>δ NPSMD SMD</td>
</tr>
<tr>
<td>Overall</td>
<td>0.26 0.38 0.47</td>
<td>0.23 0.32 0.39</td>
<td>0.46 0.78 0.85</td>
</tr>
<tr>
<td>Male</td>
<td>0.27 0.4 0.49</td>
<td>0.22 0.31 0.37</td>
<td>0.46 0.78 0.85</td>
</tr>
<tr>
<td>Female</td>
<td>0.27 0.39 0.48</td>
<td>0.24 0.34 0.42</td>
<td>0.48 0.82 0.88</td>
</tr>
<tr>
<td>Low Decile</td>
<td>0.25 0.36 0.43</td>
<td>0.23 0.33 0.42</td>
<td>0.46 0.77 0.84</td>
</tr>
<tr>
<td>Mid Decile</td>
<td>0.3 0.45 0.56</td>
<td>0.24 0.34 0.41</td>
<td>0.52 0.92 0.99</td>
</tr>
<tr>
<td>High Decile</td>
<td>0.24 0.34 0.45</td>
<td>0.23 0.32 0.39</td>
<td>0.41 0.66 0.76</td>
</tr>
<tr>
<td>NZ Euro</td>
<td>0.26 0.38 0.47</td>
<td>0.27 0.4 0.45</td>
<td>0.49 0.83 0.9</td>
</tr>
<tr>
<td>NZ Māori</td>
<td>0.27 0.39 0.49</td>
<td>0.19 0.26 0.33</td>
<td>0.44 0.72 0.8</td>
</tr>
<tr>
<td>Pasifika</td>
<td>0.28 0.42 0.5</td>
<td>0.18 0.25 0.33</td>
<td>0.44 0.72 0.8</td>
</tr>
<tr>
<td>Other</td>
<td>0.27 0.4 0.49</td>
<td>0.22 0.32 0.41</td>
<td>0.51 0.88 0.93</td>
</tr>
</tbody>
</table>

**Colour Legend**

Progress relative to other subgroups in cohort

- Lowest
- Average
- Highest

1 Note that the colours for overall effects (two years) are calibrated separately from single year effects.

### 3.4. Chapter Summary

The results presented in this chapter provide support for the assertion that decisions about which effect size metric to use and exactly how to calculate it, can have considerable impact on the resulting conclusions about the magnitude of the effect. The effect size measures utilised in this chapter primarily highlight the discrepancies that arise due to departures from normality assumptions, but other considerations, such as whether to include the correction factor and which
standard deviation to use (e.g., pooled or control; sample or assumed population) can also impact on the comparability of effect size metrics. Despite the potential for sometimes considerable differences in interpretation arising due to differences in specific calculation practices, these decisions are frequently not adequately specified in journal articles (Baguly, 2009; Ledesma et al., 2009; Leech & Onwueguzie, 2002; Peng et al., 2013; Romano et al., 2006), meaning that studies are not necessarily publishing comparable metrics (e.g., Fern & Monroe, 1996; Hogarty & Kromrey, 2001; Peng et al., 2013). This thesis does not take a particular position about which effect size measure is preferable, since this differs according to the research question and context, but suggests that since it is relatively easy to convert the various effect sizes into the metric required (see Morris & DeShon, 2002), a better focus is to ensure that researchers understand the importance of clearly describing which effect size calculation they have used – which is demonstrably not yet the case (Ahn et al., 2013; Cumming et al., 2007).

The OR analyses were of limited utility due to lack of power, with the majority of effects falling below significance. However, the patterns were interesting, are robust to violations of normality, and provide an intuitive means of comparing effects among specific subgroups. ORs are useful in cases where a specific, binary outcome is of interest (e.g., whether a student passes or fails their doctoral thesis), but make little sense when the outcome variable is continuous, since the positive features of ORs do not come close to making up for the level of information lost in reducing the scale to a simple pass or fail. Results indicated that while progress was above expectation for the majority of students whose baseline scores were in the lowest 20% of their year level, the odds of improvement were typically lower among high risk subgroups. This suggests that while LPDP appears to have been effective at lifting achievement among the majority of students with initially low scores, the project may not have been sufficiently targeted with respect to specific subgroups.

The comparison of the NPSMD based on $\delta$, against the SMD, illustrated that in all cases the SMD estimate of effect size showed inflationary bias, suggesting larger shifts than those indicated by the underlying distributions. In the writing cohorts, this bias was generally small enough that researchers would likely make the same substantive conclusions irrespective of which effect size
is calculated. However, the differences in the reading cohorts were large enough that, in some cases, conclusions would likely differ as a result of choosing to use the SMD estimate of effect size. For example, in the 2008-2009 cohort of schools with a reading focus, the overall SMD for Pasifika was 0.38; almost double the unbiased NPSMD of 0.2. In addition, the SMD effect for Pasifika was the same as for NZ European students, which would suggest that progress was equivalent for both groups, when distributional shift indicates that the shift among Pasifika was notably lower. The degree of bias was unpredictable with respect to departures from normality – although the SMD and NPSMD would be identical when comparing two true-normal distributions, larger departures only loosely related to larger discrepancies between the estimates. Instead, with respect to practical significance, the SMD was most problematic when effect sizes were small to moderate, with the inflation associated with the SMD sufficient that researchers may conclude modest gains when the distributional shift actually indicates almost no real gain. As a result, it is suggested that even when SMD assumptions are only moderately violated, a non-parametric alternative should always be investigated, especially when SMD effects are small to moderate (e.g., less than 0.5). Cliff’s delta is a useful alternative since it can easily be converted into the same scale as the SMD, but there are various other possibilities (see Hartung, Knapp & Sinha, 2011; Rogmann, 2013).

Both the NPSMD and the SMD results indicated that effects were much larger in writing than for reading, and were larger in the 2008-2009 cohort than in the previous one. These results suggest that the project may have been improving, with larger effects in subsequent iterations. In addition, there are much clearer effects for writing, with the effects in reading generally modest. It may be that baseline teacher knowledge for how to teach writing was lower, so there was more room for the facilitation to have an effect on teacher practice. In reading, the effects were also inconsistent with respect to the target groups, with NZ Māori students, and those attending low decile schools gaining less than other students in the first reading cohort. In contrast, in the second reading cohort, the SMD results suggest that the target groups gained at about the same rate as for the project overall, while the NPSMD suggests that Pasifika gained less than other students. In writing, the gains among students in low decile schools were equivalent or better
than the project overall for both cohorts. Pasifika students progressed marginally more rapidly than other students in the first writing cohort, while Māori and Pasifika both gained marginally less than other students in the second writing cohort. These results suggest that the project was more successful in eliciting shifts in writing progress, but that there was minimal differential effectiveness. That is, students in the target groups did not consistently gain more than other students – nor did the gap increase, however.
CHAPTER 4. SINGLE-LEVEL REGRESSION

In Chapter 3, various effect size measures were introduced, and the results of LPDP analysed using a selection of these metrics. There were differences in the magnitude of the effects depending on the procedure chosen, but irrespective of method, the gains described by the effects were large for students in schools focusing on writing and more modest for those focusing on reading. The current chapter again makes use of the LPDP data to investigate the research questions of the thesis, in the current case by using the concepts of regression analysis to build different kinds of models. Initially, the analysis incorporates only one additional factor compared with the effect size analyses in Chapter 3; that of the distance of students’ baseline scores from the normative mean in order to estimate the effect of regression to the mean. Subsequently, more complex models are investigated, incorporating all factors that were identified to have an effect on the score at the end of the second year of the project, and for which data were available. These first two sets of analyses provide the main contribution to the primary aim considered within this chapter, and also provide results that are easily compared to the results in Chapter 3 since it is possible to convert these to the same unit of measurement (standard deviation units, or SDU). The third set of analyses use value-added models to determine whether there were differences in the relative contributions of individual facilitators. These analyses do not directly address the overarching research question within the current chapter, but do provide insight into project effectiveness, which is the focus of the secondary research questions. In addition, these analyses will be revisited in Chapter 5 using hierarchical linear modelling, allowing for a comparison of the results to determine the extent to which conclusions about the facilitator effect differ as a result of the different methods, thereby addressing the primary aim of the thesis.

A more detailed description of the structure of Chapter 4 is as follows. The first section of this chapter provides a general review of the regression concepts (section 4.1) that support the analyses conducted in the analysis section (section 4.2). The first subsection of this review briefly discusses the historical origins of regression, particularly due to its origins in understanding
regression to the mean, which is an important consideration for educational data, particularly where group effects of students that differ from the population mean are considered. A more general review of regression is included in subsection 4.1.2, introducing the basic regression equation on which the analyses are based. The steps discussed in subsection 4.1.3 explain the process of regression analysis in more detail. These steps are intended to be generalizable to other contexts, but also describe the process used in the analyses conducted in this thesis. Particular focus is given to the concepts required to understand the procedures used in the analysis section. The analysis section follows a similar order, with progress initially considered with respect to regression to the mean (section 4.2.1) followed by development of more ‘inclusive’ regression models in which several factors are considered concurrently (section 4.2.2). The final subsection of the analysis section presents the value-added models indicated in the previous paragraph. These models are an extension of the ‘inclusive’ models developed in the previous subsection, and investigate whether there are differences in the rates of progress among students attending schools that were working with different facilitators. The final section of the chapter summarises the major findings of the chapter. Again, particular focus is given to differences in conclusions about project effectiveness that result from the different methods. This focuses primarily on differences that arise within this chapter, with differences between chapters addressed in Chapter 6, but findings of particular significance are signposted nonetheless.

4.1. Introduction to the regression concepts utilised in this chapter

This section introduces the basic concepts and statistics that are needed to understand the results of the regression analyses presented in section 4.2. Within the constraints of the thesis structure this cannot be an exhaustive introduction, but the concepts required to understand the analyses developed within this chapter are explained, with particular focus on linear regression. For a more complete discussion, see “Regression Analysis by Example” by Samprit, Ali and Bertram, (2006), which has been a valuable resource in the development of this Chapter.
4.1.1. Regression to the mean

The notion of “regression” was initially conceptualised by Francis Galton more than 100 years ago, to describe the biological tendency for children of particularly short or tall parents to regress toward more average heights (Galton, reprinted 1989). Despite a strong association, or correlation, between parental height and the height of their children, it was not possible to predict the height of future offspring precisely, with the differences between actual and predicted heights trending toward the population mean. This phenomenon is described as “regression toward the mean” (RTM), since measurements tend to be closer to the average on subsequent measurements (Everitt, 2002; Kelly & Price, 2005; Stigler, 1997; Upton & Cook, 2006). RTM occurs for all measurements above or below the mean, but as the initial distance from the mean increases, so does the subsequent regression toward the mean (Kelly & Price, 2005; Lohman & Korb, 2006; Marsh & Hau, 2002).

Although measurements that are further from the mean at baseline will show greater RTM than those around the mean, the degree of regression to the mean also depends on how reliably the predictor and outcome variables can be measured; that is, the degree of random error inherent in the measurements (Sheskin, 2003). Measurement error is introduced by a number of factors, some of which are systematic in nature, while others are non-systematic. Systematic bias is caused by factors that affect test scores in a methodical way, for example, if the test consistently yields results that reflect systematically higher achievement than students’ underlying ‘true’ ability. Systematic errors neither cause nor prevent RTM however, and can generally be mitigated via appropriate research and test design (Barnett et al., 2005), so are not discussed in detail here. Non-systematic errors are more difficult to prevent since they are inherently random, but there are a number of different approaches that can mitigate the magnitude of the effect at the design stage, or attempt to account for or quantify the RTM effect at the analysis stage.

Using Galton’s height example, if one measures the same individual’s height twice in quick succession, one would expect a correlation of almost 1.0 between the two measurements since the heights should remain almost constant between two closely spaced measurements, and height can be measured with a high degree of measurement precision. Despite this, there would
be a degree of random variability around each individual’s true height caused by various non-systematic factors, such as minor differences in posture at each measurement, or minor variations in the angle of the ruler used and so on. Conversely, a source of systematic error might relate to the particular choice of measurement (e.g., a wooden metre ruler that has become slightly bent), or if we were collecting heights at a number of different testing sites, each site might use slightly different methods. As noted above, systematic errors do not cause RTM effects (Barnett et al., 2005), but are important to consider since they are another possible threat to the validity of our conclusions, and are primarily introduced or mitigated by research design (Jusko, 2008). For example, in the LPDP sample it was important to ensure that the tool was reliable and valid as discussed in Chapter 2, and the involvement of facilitators helped to ensure consistency of test administration and marking across schools and over time.

Non-systematic, or random errors, are invariably present in educational measurements. As noted, the impact of random error is reduced when measurements can be obtained with a high degree of precision, so test and research design are important, but in the educational context we rarely have the luxury of precise measures; tests are imperfect estimates of students’ abilities affected by a huge range of factors that contribute to the imprecision of these estimates (Amrein-Beardsley & Barnett, 2012). A single test result only gives indicative information about a student’s ability in the test subject; assessments cannot provide a perfect measurement of a student’s ability (Rudner & Schafer, 2002). For example, inadequacies in test design, student fatigue, hunger, luck and home factors all affect our precision when attempting to determine an individual’s ability, and the non-systematic aspects of these factors are extremely difficult to predict or account for (Smith & Smith, 2005). Since there are so many contributors to error in educational measurements, the tendency for measurements to regress toward the mean is especially strong in the educational setting. It is also important to note that since RTM is caused by non-systematic, random error, producing variability of each observed measurement around a theoretical “true” mean for that individual, it has a greater effect at the distributional extremes. This is because such individuals are more likely to have been especially “lucky” (or unlucky) at baseline than those with achievement around the mean. In addition, the RTM effect describes a
group level effect, meaning that although certain individuals might obtain equally extreme (or even more extreme) measurements, at a group level the effects of variability invariably show a tendency for initially extreme measurements to become less extreme upon subsequent measurement (Barnett et al., 2005).

Since the RTM effect has an increased effect the further the measurement is from the mean, consideration of this phenomenon is especially important for educational interventions focusing on improvements for students with especially low (or high) achievement (Barnett et al., 2005). The phenomenon is correlational in nature, meaning that in a typical classroom, the lowest achieving student will not necessarily score better in subsequent tests but, on average, students with the lowest scores will improve since the effects of chance are stronger for those at the distributional extremes (Randolph, 2008). Analyses investigating the progress of particular subgroups (e.g., splits by ethnicity, lowest 20%) are especially susceptible to RTM effects; since the group mean differs substantially from the population mean, there will be a greater degree of regression toward the mean evident in subsequent measurements for that group. This has particular relevance to the educational context, since policy often focuses on target groups for whom RTM effects are frequently more pronounced. In addition, although extreme measurements trend toward the mean as a group, the second set of measurements is likely to show the same variance and mean as the first (Kelly & Price, 2005). This seems counter-intuitive since RTM predicts a tendency for individuals at the extremes to regress toward the mean, implying that the range (and therefore variance) of scores will decrease. However, since RTM is caused by the extent to which the measurements are due to chance, the variability around the mean tends to remain the same unless there are differential effects for particular subgroups (Kelly & Price, 2005). This is because each measurement is obtained in the presence of additional error effects, meaning that each measurement will show variability around the score predicted by the RTM effect, resulting in equivalent variance unless there is a differential effect over and above that predicted by RTM (Barnett et al., 2005). The degree to which the RTM effect affects the conclusions made about the LPDP sample is discussed in the analysis section of this chapter in subsection 4.2.1.
4.1.2. Regression analysis more generally

Although the term ‘regression’ heralds from its conceptual origins as a way of describing the regression to the mean effect (which remains an important consideration for educational research), regression analysis has become popular for analyses well beyond the RTM scope. This popularity is largely because of the predictive capabilities of regression demonstrated in early work by Pearson and Fisher. This work demonstrated that the degree to which a variable or outcome (the dependent variable), is affected by other, explanatory variables (independent variables), could be formalised as a regression equation, then used to estimate or predict the most likely outcome variable values based on the known information derived from the independent variables. In addition, regression analysis is conceptually appealing since it allows for prediction of real-world outcomes in an intuitive way (Gorard, 2012). For example, extending Galton’s original height example, knowledge about parents’ heights would be useful if one were attempting to predict the height of their children. Without any additional information, it would be reasonable to predict that the children of two very tall parents would likely be taller than the children of a very short couple. However, it is important to understand that although regression equations attempt to formalise the relationship between the variables by minimising the prediction error, as long as error exists, the regression equation is simply a best prediction based on the information available. Continuing with the heights example, since the association between the heights of parents and their children is not perfect, it should be clear that individual predictions are unlikely to be perfect. It may well be that the children of the short parents in this example turn out to be taller than those of the tall parents – perhaps because the children of the short parents are male while the others are female, or simply due to chance. Since we know that gender is also associated with height, this additional information would be useful for developing a more accurate prediction. This demonstrates, at a very basic level, the way in which regression analysis can utilise available information to move beyond pairwise comparisons such as the effect sizes presented in Chapter 3, to calculate the best prediction of an outcome, as well as illustrate the extent to which each independent variable has an impact on the outcome.
A simple univariate (only one response variable) regression model of a random subset of the 2008-2009 Year 7 LPDP reading data, using baseline achievement to predict subsequent achievement the end of the first year of the project (EOY1), is shown in Figure 5. Only 10% of Year 7 students are included since this is intended to be illustrative only, and inclusion of all students would make it very difficult to see individual points in the scatter plot. The regression equation in this example follows the generic linear equation form, \( y = mx + c \); where \( y \) is the predicted value, \( m \) is the slope of the line or degree to which \( x \) has an impact on \( y \), while \( c \) is the point at which the line intercepts the \( y \)-axis. In this case, since the asTTle minima are equal to 100, the \( y \)-intercept is not inherently meaningful, and simply allows the line to be positioned in the location of best fit. The \( R^2 \) value of .47 indicates that almost half of the variability in EOY1 scores is explained by students' time one scores – not a terribly unusual result. It also indicates that the relationship between baseline and EOY1 scores is not perfect. There is a degree of scatter around the line, which reflects the degree of prediction error inherent in the regression equation. Most likely, additional precision could be obtained using information from other covariates, but some scatter will remain due to measurement error.
Figure 5. Basic regression equation showing the average relationship between baseline and EOY1 asTTle reading scores for a 10% sample of Year 7 students in the 2008-2009 cohort.

Since the slope (m) is less than one (.65), this indicates that students with lower initial scores gained more than those with higher initial scores. Using the equation, this would mean that the predicted time two score for a student with a baseline score at the 20th per centile for this group would be 501 \((y = 447 \times 0.65 + 210.7)\), constituting a gain of more than 50 points over the initial score of 447. For a student in the 80th per centile, the predicted gain would be just ten points, from a score of 573 at time one to a score of 583 at time two \((y = 573 \times 0.65 + 210.7)\). This relationship is likely to be at least partly explained by regression to the mean, which will be quantified formally in the analysis section investigating the RTM effect. It is important to note that this example is intended to be illustrative only.

4.1.3. Steps in a regression analysis

In order to conduct a regression analysis there are a number of required steps, including:

- Statement of the research question
- Selection of potentially relevant variables
- Data Collection
- Model Specification
These steps are adapted from Samprit and colleagues (2006) and are discussed briefly below, since these steps were followed in the analyses conducted in section 4.2.

4.1.3.1. Statement of the research question(s)

This step may seem out of place in the sense that the research question is typically situated as an issue of research design rather than analysis per se, but meaningful analysis is inherently reliant on adequate preparatory work at the research design stage. A well-constructed research question allows the researcher to tailor the research methods around what is most appropriate for the research question. Conversely, an ill-posed research question can lead a researcher to select the wrong variables or predictors, choose an incorrect model or analytic methodology, and ultimately, lead to a waste of considerable time and money. While financial constraints were less of an issue in the current context since the data collection phase had already been completed, it was important that the research questions be appropriately operationalized to ensure that the analyses adequately addressed these questions.

4.1.3.2. Selection of potentially relevant variables

This step must be theory-driven, and should take place before any data collection. This step also includes decisions around how to measure the particular variables that have been identified as relevant to the research question(s) based on the literature. If for example, one is trying to determine the factors that influence reading progress (this was one of the LPDP research questions), the researcher(s) would need not only to identify factors that have previously been identified as likely to have an impact on reading progress (e.g., socio-economic status), s/he would also need to ascertain an appropriate way of measuring these factors quantitatively. First,
an appropriate tool for measuring the dependent variable should be identified (in the LPDP sample, this was the asTTle tool), and the measurements from this tool would be used as the response variable data. At this step, it is also important to be aware of the analyses that will likely be conducted with the data as many analyses have different requirements. For example, as noted in section 2.4.1.2, the PAT (Reid & Elley, 1991) data collected during the LPDP project were not considered in these analyses since the resultant stanine scores are non-continuous and have significant ceiling effects, seriously limiting the range of possible statistical applications that could be explored. The decisions around how to measure the response variable are arguably the most important since all other variables will be tested in relation to this. Irrespective of relative importance however, similar decisions must be made about how to measure adequately the x-axis covariates (the independent or predictor variables) after identifying which variables should be measured based on the literature.

4.1.3.3. Data collection

There are a number of issues that must be overcome within this step, to ensure that the regression analysis is based on robust data. Typically in educational contexts the researcher has limited control over the factors that are not of primary interest in the research, but this can be mitigated by collecting data about these factors, or by designing the research in a way that reduces the impact of such factors. For example, much of the demographic information that is collected may not be directly related to the research question, but allows the researcher to estimate and control for differences among participant subgroups. This information can be used prospectively to assign participants to specific equivalent groups or, retrospectively, to quantify differences between groups. The LPDP data collection included several demographic variables, such as ethnicity and school decile rating. All of the demographic information collected was of some interest to the research questions, specifically with respect to the extent to which there were differential gains among target subgroups.
Participant attrition and missing data also need to be minimised as much as possible at this stage. Many of the regression algorithms will exclude participants with missing data in a listwise manner, meaning that a single missing datum results in the entire set of data for that participant being excluded. There are various approaches for quantifying and reducing the bias that arises from missingness, but ultimately minimising this at the data collection phase is the ideal. Attrition was a major consideration within the LPDP sample, with considerable loss of students in the second year of the project. Imputation was unsuitable since students with missing test scores had only one or two other assessment instances, as discussed in section 2.1.5. In addition, the proportion of students with missing demographic information was negligible.

4.1.3.4. Model specification

Model specification involves the researcher using a combination of existing knowledge, as well as both objective and subjective judgments to make expert decisions about what form the model should take (Samprit et al., 2006). More specifically, this argument indicates that the model specification is not fixed; rather, it is a process whereby the researcher determines which of the variables gathered in the data collection phase should be retained in the model being tested. The model will already have been conceptualised during the research design phase, but it is not always clear what form the data will take prior to its collection, nor the specific difficulties that may arise, nor the relationships that will prove predictive of outcomes. As a result, the conceptualised model may need to be iteratively updated to incorporate this information after preliminary analysis.

As previously indicated, the regression model attempts to estimate the true relationship between one or more predictors using the aggregate information of these predictors, and one or more outcome variables. Most commonly, models developed in education and the social sciences are multiple regression models; where two or more predictor variables are used to estimate a single outcome variable as shown in the following equation:

\[ Y = f(X_1, X_2, \ldots, X_K) + \epsilon \] (7)
This equation indicates that the outcome variable, \( Y \), is a function of a set of \( X \)-variables, where \( K \) specifies the number of predictors and \( \varepsilon \) is a residual term that accounts for the random error that arises in the model when the relationship is not perfect (i.e., there is scatter due to chance, error and other relevant variables being omitted or not measured – that is, always!), resulting in an inability for the regression equation to provide a perfect model fit. The coefficients derived from the regression analysis can then be used as an estimate of the magnitude of the effect of the predictors on the outcome variable, or generalised to predict outcomes in another cohort. Since regression analysis is correlational in nature, causality is difficult to determine without additional analysis and / or design constraints.

### 4.1.3.5. Selection of fitting method

Typically the researcher should have a clear idea about which predictors to include in the regression model, based on previous literature. However, in some cases, there may not be much guidance available from past research, or it might be that a number of relevant variables are suggested by the literature. As a result, a number of different procedures have been developed to guide variable selection via statistical procedures allowing the predictors to be entered into the regression equation in different ways. The default procedure is simply called the enter technique, whereby all variables are entered into the regression equation simultaneously, without regard to any particular order or relative importance (though there is a minimum tolerance), and the model simply tests how effective these predictors are at predicting the response variable. This procedure is therefore highly reliant on the researcher ensuring the relevance of the predictor variables being entered into the regression equation.

Other techniques of variable selection are stepwise procedures, whereby each variable is added to the regression equation iteratively in different orders. These techniques are sometimes called hierarchical regression, though this is becoming less common and is avoided here due to the possibility for confusion with hierarchical linear modelling which is discussed in Chapter 5. The forward selection procedure initially fits a regression equation with just a constant term and
no predictors. This first step in the forward selection procedure yields the overall grand mean of the outcome variable since no other predictors are included. Regression equations are then fitted with one additional predictor at a time, with variables added in the order of the magnitude of the simple correlation with the response variable. Backward elimination starts with the full equation (that is, the same as the enter procedure), but then eliminates one variable at a time based on how much the variable reduces the error sum of squares (discussed in more detail later in this subsection), in the order of degree of error reduction from smallest to largest. Both forward and backward selection utilise significance criteria, but since the definition of significance is flexible some algorithms will simply continue deleting variables until only the equation with the constant remains, allowing the researcher to determine which variables ought to be retained. Full stepwise regression is the most computationally sophisticated method, and is effectively a combination of both forward selection and backward elimination. The model fitting procedure follows the forward selection process, but allows backward elimination at each step. This means that although predictors are added in order of the strength of their relationship to the response variable, they are also tested after the addition of each additional variable to ensure that the previously added predictor(s) continue to contribute to the predictive ability of the model (Brace, Kemp, & Snelgar, 2009).

It should be noted that while stepwise procedures tend to produce more parsimonious models, such models are often criticised as being analogous to data dredging due to the possibility of over-fitting of a model since these procedures are more sensitive to sampling error variations than the enter method. As a result, apparent relationships may only be valid within the specific data set, resulting in limited generalizability though these issues can arguably be overcome with large datasets using split-half or two studies (Copas, 1983; Rencher & Pun, 1980). However, despite considerable discussion about the inadvisability of stepwise methods, stepwise regression remains extremely popular in education (Onwuegbuzie & Daniel, 2003), so these methods are used in the analysis section of this chapter as an additional point of comparison with the HLM analyses in Chapter 5 (since these models do not allow stepwise methods).
Following selection of the fitting method, the way in which the predictor is related to the response variable must be determined. Again, previous research will guide this, but bivariate scatter plot graphs showing the relationship between each predictor and the outcome variable are extremely useful for providing an indication of how the variables are related to each other. If the nature of the relationship is unclear, it is possible to test a number of different relationships to determine which describes the relationship best. Some common examples are shown in Figure 6 below using the same data described in section 4.1.2; that is, 10% of Year 7 students attending schools with a reading focus in the 2008-2009 cohort. The $R^2$ value provides a reasonable indication of how well the equation describes the relationship between the two variables. However, $R^2$ tends to increase with any added complexity, whether in the form of additional variables or allowing additional complexity (nonlinearity) in the relationship between the predictor and response variables and, as a result, is argued to overestimate the explained variance with respect to future samples (e.g., Carter, 1979; Fan, 2001; Snyder & Lawson, 1993; Thompson, 1999; Yin & Fan, 2001). This can be seen in the examples shown in Figure 6, where the quadratic, cubic and quartic equations have incrementally higher $R^2$ values, yet do not visibly provide a better description of the data – indeed, most of the distinction between the equations is seen in the distributional extremes, where there are fewer data points and the position of each datum is less reliable.
Figure 6. Various regression equations describing the relationship between baseline and EOY1 asTTle reading scores for a 10% sample of Year 7 students in the 2008-2009 cohort of LPDP.

As a result, R2 values are frequently adjusted to penalise the added model complexity using variations of an equation that penalises small sample size and model complexity. The most commonly used method is the Ezekiel correction which is shown below:

\[
R^2_{\text{adjusted}} = 1 - (1 - R^2) \left( \frac{n - 1}{n - d - 1} \right)
\]

Where: \( n \) is the number of cases, and \( d \) is the order of the polynomial.

Based on this equation, the adjusted \( R^2 \) values would all decrease marginally compared with those shown in Figure 6. The \( R^2_{\text{adjusted}} \) would be highest for the cubic equation, at 0.4717, but despite this, the linear case, with an \( R^2_{\text{adjusted}} \) of 0.4676, would still provide the best fit for the data. The gain in \( R^2 \) is rather minimal, and is solely because the higher order polynomials are more responsive to outliers, which is undesirable since it limits generalizability and places too much confidence in the measurement precision of the data. Although the Ezekiel method is considered

---

4 Note that this adjustment of the \( R^2 \) value follows the Ezekiel method, but there are a range of different equations in the literature. All of these penalise small sample sizes and model complexity to varying degrees. For a full discussion of the various adjustment methods, see Leach and Henson (2007).

5 Linear terms are first order polynomials so \( d \) would equal one in the linear case, two for a quadratic and so on.
a conservative estimate of $R^2$, the adjustment is based on a population effect. Since educational researchers frequently intend to draw conclusions about the effects of certain predictors on an outcome for policy decisions or some other generalised, future sample, the adjustment should actually be even more conservative (Leach & Henson, 2007). Regardless, the Ezekiel method has been used throughout this thesis for pragmatic reasons: adjusted $R^2$ values reported by SAS and SPSS use the Ezekiel method (Kirk, 1996), so values using this method are more easily compared with previous research; and the correction made by all of the available formulae gets smaller as sample size increases, so the difference between individual methods is negligible in the LPDP sample. Where the adjusted $R^2$ indicates only a modest improvement of model fit from added model complexity, the linear approximation should be used regardless, since it provides minimal loss of precision and more easily interpretable results (Leach & Henson, 2007).

Once the researcher has determined the form of the model (what predictors to enter and for what outcome), the next step is to estimate the model parameters based on the available data. Determining the parameter estimates that provide the best description of the available data is essentially an optimisation problem, with various options available in most statistical packages, and the researcher must determine which is most appropriate. Regression models are classified as linear or non-linear based on the mathematical function of the parameters (that is, the coefficients being estimated) in the equation, rather than the relationship between the $X$- and $Y$-variables. This is because nonlinear functions of $X$ can be re-expressed as linear functions, while cases where the parameter enters the equation nonlinearly cannot. As a result, although nonlinear relationships between the predictors and outcome variables were considered in the analyses undertaken as part of this thesis, all of the parameter estimates enter the regression models linearly, so nonlinear regression is not discussed in detail. Nonlinear regression is comparatively rarely used in educational applications, but for a comprehensive discussion see Bates and Watts (1988). The generic form of the linear regression model is shown below

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \varepsilon$$  (9)
Where: the beta-coefficients are the parameters being estimated, with \( \beta_0 \) representing the intercept, and the other betas representing the slopes of the lines for each predictor. The \( \epsilon \) represents the predictive errors, which are assumed to be random, and normally distributed with a mean of zero.

The parameters in linear regression models can be estimated using several different methods, with each method optimising slightly different aspects of the model and minimising error in different ways. Most regression models are estimated using a form of the least squares method, in which the sum of the squared errors is minimised. Error in a regression model can be defined in a number of ways, but the simplest and most commonly used is the ordinary least squares method, which minimises the sum of squares of the vertical error (or residual). The vertical residual is the difference between the predicted outcome derived from the regression model and the actual outcome measured during the data collection phase. This is the most appropriate form of error minimisation for the analyses used in this thesis since this minimises the error in the response variable, providing parameter coefficients that are optimised with respect to predictive accuracy of the asTTle score.

Other least squares methods, including orthogonal regression and partial and weighted least squares, minimise the sum of squared errors by minimising or defining error differently, and are often used to attempt to correct for violations of the assumptions required to make the ordinary least squares estimates valid. For example, orthogonal regression minimises the perpendicular error – that is, the regression parameters returned provide a trend line that simply minimises error without regard to a particular axis (the error is measured as the distance perpendicular to the regression slope). This method is typically used to correct for measurement error in the predictor variables, but tends to overcorrect since equation error is ignored (Carroll & Rupert, 1996).

### 4.1.3.6. Model fitting

Once the fitting method has been selected, the actual regression analysis can be conducted by fitting the model to the available data using the selected method of estimation. As indicated
above, the most appropriate estimation method for the data analysed throughout this thesis appeared to be the ordinary least squares technique, which is available in most statistical software packages, though this was still assessed during the model validation step. After determining the likely predictors for the response variable of interest, the parameter estimation and model fitting are automated processes offered by a wide range of statistical software, including SPSS⁶, SAS⁷ and R⁸. The researcher must select what variables to include in the model, the method of estimation, and how the predictors are entered into the model fitting process, but the software provides the calculations of the parameter estimates based on these constraints.

4.1.3.7. Model validation and critique

Validation of a regression model is an iterative process in which the fitted model must be examined with respect to statistical assumptions and theoretical interpretations, and then used to determine the performance and applicability of the model and guide any modification of earlier steps in the analysis (e.g., addition or removal of predictors, data transformation, estimation techniques and so on). The assumptions made while determining the form of the model (i.e., which predictors, estimation method, linearity and so on) must hold if the model is to be valid, so adequate planning and assessment of the form the model should take helps to expedite the validation process.

The simplest way to test whether model assumptions hold is to assess the error distribution, particularly via examination of residual plots. The errors (also called residuals) are assumed to be independently and identically distributed (iid) with a mean of zero and variance of $\sigma^2$. In other words, the residuals should follow a normal distribution and be uncorrelated. This can be assessed using a residual plot, or via the Kolmogorov-Smirnov test to determine whether the standardised residuals follow a normal distribution with a mean of zero. The standardised residuals should also have minimal correlation with each of the predictors, with a scatter plot of

⁷ http://www.sas.com/
⁸ http://www.r-project.org/
the residuals against the predictors showing no detectable patterns. Graphical methods are considered particularly useful for confirming the performance and validity of regression models since graphs can provide insight into features of the data that are extremely difficult to identify using formal diagnostics (Huber, 1991). These assumptions were tested with respect to the regression models developed in section 4.2.2. Since these models formed the basis of the analyses presented in the value-added section (4.2.3), it was of particular importance that these assumptions be checked.

4.1.3.8. Usage and reporting of the derived model

Ultimately the intention of regression analysis is to present a model that describes the relationship between a series of predictors and a particular outcome of interest. This can then be used to inform policy, measure the effects of changes in policy, assess relationships and inter-relationships between variables, or predict likely outcomes for individual students and use these predictions to provide targeted teaching. In order to ensure that the conclusions drawn from the developed models remain valid, it is essential that the regression assumptions be assessed and that conclusions and predictions remain within the range of data that made up the original model.

In the current context, it was intended that the regression models would build on the effect size analyses presented in Chapter 3, allowing for a more nuanced interpretation of the shifts elicited by LPDP. Since the regression models partial out the variance explained by covarying factors such as decile and ethnicity, it should be expected that there would be differences in the results of these analyses, compared with the effect size analyses. However, as noted in the previous chapter, meta-analyses conducted in education frequently exhibit a lack of information with respect to original study methodology, meaning that some of the effect sizes are likely to be pairwise effect sizes, while others take into account other covarying factors (Peng et al., 2013). This is likely to result in inflated effect sizes for studies reporting the former, compared with those that build more nuanced models. Differences in the results presented in this chapter for the LPDP data, compared with the effect size analyses presented in Chapter 3, allow the extent to which
such decisions result in different conclusions about these data to be assessed. In addition, these analyses assist with building a more complete picture about progress among students in LPDP, helping to determine the extent to which the project was successful in accelerating progress for priority subgroups, and for students overall.

4.2. Analysis

This section utilises single level regression procedures to investigate the pattern of progress within the four LPDP clusters utilised throughout this thesis (i.e., the two reading and two writing cohorts, from 2006-2007 and 2008-2009). The effect of regression to the mean (RTM) is addressed initially, since an understanding of this effect helps to contextualise patterns identified (section 4.2.1). The size of the RTM effect differs depending on where the group lies with respect to the remainder of the distribution, so the effect is quantified among each of the subgroups investigated throughout the thesis – specifically, by gender, ethnicity and school decile. In the following subsection (4.2.2), a series of regression models are developed in order to investigate the effect of being in one of the subgroups identified above, on achievement and progress. The results of this subsection are the most important part of this chapter in terms of the primary aim of the thesis, since these results are more readily comparable with the results of the other analyses. The effects of several other variables are investigated; the results of which assist with conclusions about project effectiveness. Finally, subsection 4.2.3 uses value-added models to investigate certain higher-level effects, such as school- and facilitator-effects. The models used in this section are derived from those determined to provide the best fit for the data during the analyses undertaken in subsection 4.2.2. The results of the value-added analyses offer a precursor to the more complex hierarchical models presented in Chapter 5.

As with the analyses in the previous chapter, asTTle scores were standardised across year levels to allow for aggregation into four separate cohorts of students, by subtracting the expected asTTle score for each student’s year level based on the asTTle normative sample. This facilitates interpretation of relative achievement levels by indicating the extent to which students are
achieving above or below expectation, and reduces complexity by mitigating the effect of year
level as a specific variable.

4.2.1. Regression to the mean

Although there have been many articles over the last several decades discussing the effect of
regression to the mean (RTM), combined with considerable commentary about how the RTM
effect can cloud the interpretation of change within longitudinal studies, attempts to quantify the
RTM effect precisely are confounded by various definitional issues. As noted in section 4.1.1, the
RTM effect is a function of the distance from the mean, and the correlation between the baseline
and follow-up scores. When the measurement is at the mean or the two measurements are
perfectly correlated, there will be no RTM effect. Smith & Smith (2005) formalise the RTM effect
in the equation below:

$$\hat{\mu} = (1 - \rho^2) \bar{X} + \rho^2 \bar{X}$$

(10)

Where $\hat{\mu}$ is the predicted score, given the distance of the individual (or group) from the
population mean, adjusted for $\rho^2$, defined as the reliability of the test (Smith & Smith, 2005). This
equation is derived from the ordinary least squares procedure discussed in section 4.1.2 allowing
the RTM effect to be separated out from intervention effects and other independent variables.
The difficulty with the application of this equation however, is that the population mean and the
reliability can both be difficult to define. For example, if one is assessing the RTM effect among
boys in the LPDP sample, should the population mean be defined as the overall asTTle
normative mean, or the normative mean specifically for boys? Or should the population of schools
from which the LPDP sample has been drawn be expected to have a lower mean – this seems
reasonable since schools in the project had self-identified as needing additional support, which
would indicate that the population mean for the LPDP cohorts might, in fact, be lower than the
mean indicated by the normative sample. In addition, there are various methods reported in the
literature for deriving the reliability, including derivations based on Pearson’s correlation
coefficient (Pearson \( r \)), the intra-class correlation coefficient (ICC) and limits of agreement (LOA) (Weir, 2005), further complicating attempts to quantify the RTM effect.

In the current case, the procedures outlined by Smith and Smith (2005) have been followed, with reliability \( (\rho^2) \) defined as equal to the (unsquared) Pearson correlation of scores on comparable tests (X and Y), when the standard deviation of the two tests is equal. While this differs from the usual definition of test reliability, the correlation takes into account the fact that even with a perfectly reliable test, there will be regression to the mean due to other factors (such as true changes in underlying ability over time). The asTTle test has a standard deviation of 100 for students of all schooling year levels, by design. The mean to which scores regress has been defined as the asTTle normative mean. Although it is likely that this overestimates the RTM effect in some cases, this provides the greatest flexibility and interpretability across the four cohorts, and provides a conservative indication of the degree of value-added by LPDP for specific subgroups, after accounting for baseline score. The overall RTM effect from baseline to EOY2 is quantified against the effect sizes calculated in Chapter 3 since this provides a directly interpretable assessment of the extent to which the RTM effect has contributed to the apparent success of subgroups within LPDP. Since the asTTle scores of students within the project have already been “centred” around the asTTle normative mean for each student’s year level, the RTM equation becomes somewhat simplified (since \( \bar{X} \) becomes 0), resulting in the predicted score equalling \( \rho^2 \) (which is equal to \( r \)) times the actual baseline score.

The RTM effect is quantified using standard deviation units (SDU) to assist with interpretation and contextualisation of the effect. This is on the same scale as Cohen’s \( d \). As with the tables in Chapter 3, the tables presented describing the RTM effect use conditional formatting. The colours are the same, with, in a general sense, red indicating a poor result, yellow a moderate one, and green a positive result. More specifically, in the current section, red is used for a value-added SDU (i.e., Cohen’s \( d \)) of 0, yellow for an SDU of 0.4 (i.e., the threshold suggested to be about average by Hattie, 2009), while green is used for a value-added effect of 0.8 or greater.
4.2.1.1. Reading Focus – both cohorts

Table 15 provides an overview of the RTM effect within the reading cohorts of LPDP. The correlation between baseline and follow-up was 0.62 for both cohorts. Typically the predicted effect of regression to the mean is relatively small since the baseline scores of students within the reading cohorts were relatively close to the mean. Regression was also generally negative since many subgroups had baseline means above the estimated population mean (i.e., the asTTle normative mean). Typically the added value of the project, after controlling for the estimated RTM effect, was larger in the 2008-2009 cohort than in the 2006-2007 cohort, following the same pattern as the raw SMD effect sizes in Chapter 3. However, in both cohorts, controlling for the RTM effect resulted in quite different interpretations about the relative effectiveness of the project among specific subgroups, with groups that were usually already advantaged showing larger added value effects. For example, in both cohorts, students in high decile schools made nominally similar progress compared with students in other decile groups (indicated by the SDU), but after controlling for the RTM effect it appeared that students in high decile schools had larger added value from the project than students in low or mid decile schools. The same pattern is evident by gender and ethnicity, with female students and NZ Europeans typically estimated to have had a greater added value effect than other students.
Table 15. LPDP added value within reading by subgroup after removing estimated RTM effect.

<table>
<thead>
<tr>
<th></th>
<th>Reading 2006-2007</th>
<th></th>
<th>Reading 2008-2009</th>
<th></th>
<th>Added value of the project</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Predicted RTM effect</td>
<td>Progress (SDU)</td>
<td>Added value of the project</td>
<td>Predicted RTM effect</td>
<td>Progress (SDU)</td>
</tr>
<tr>
<td>Overall</td>
<td>-0.12</td>
<td>0.24</td>
<td>0.34</td>
<td>0.01</td>
<td>0.42</td>
</tr>
<tr>
<td>Male</td>
<td>-0.08</td>
<td>0.24</td>
<td>0.31</td>
<td>0.06</td>
<td>0.42</td>
</tr>
<tr>
<td>Female</td>
<td>-0.16</td>
<td>0.24</td>
<td>0.37</td>
<td>-0.04</td>
<td>0.44</td>
</tr>
<tr>
<td>Low Decile</td>
<td>-0.02</td>
<td>0.16</td>
<td>0.17</td>
<td>0.13</td>
<td>0.45</td>
</tr>
<tr>
<td>Mid Decile</td>
<td>-0.08</td>
<td>0.3</td>
<td>0.36</td>
<td>-0.02</td>
<td>0.36</td>
</tr>
<tr>
<td>High Decile</td>
<td>-0.23</td>
<td>0.24</td>
<td>0.42</td>
<td>-0.16</td>
<td>0.52</td>
</tr>
<tr>
<td>NZ European</td>
<td>-0.17</td>
<td>0.26</td>
<td>0.41</td>
<td>-0.09</td>
<td>0.38</td>
</tr>
<tr>
<td>NZ Māori</td>
<td>0.02</td>
<td>0.13</td>
<td>0.1</td>
<td>0.14</td>
<td>0.43</td>
</tr>
<tr>
<td>Pasifika</td>
<td>0.08</td>
<td>0.4</td>
<td>0.23</td>
<td>0.13</td>
<td>0.38</td>
</tr>
<tr>
<td>Other</td>
<td>-0.16</td>
<td>0.3</td>
<td>0.41</td>
<td>-0.01</td>
<td>0.71</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Colour</th>
<th>Estimated added value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Legend</td>
<td>0 SDU</td>
</tr>
<tr>
<td></td>
<td>+0.4SDU</td>
</tr>
<tr>
<td></td>
<td>+0.8SDU</td>
</tr>
</tbody>
</table>

4.2.1.2. Writing Focus – both cohorts

Table 16 describes the RTM effect within the LPDP writing cohorts. The correlation between baseline and follow-up was 0.58 for both of the writing cohorts. Typically the predicted effect of regression to the mean was large since the baseline scores of students within the writing cohorts were considerably below the mean of the asTTle normative sample. The regression effect was positive since all subgroups had baseline means below the estimated population mean. As with the reading cohorts, the added value of the project, after controlling for the estimated RTM effect, was larger in the 2008-2009 cohort than in the 2006-2007 cohort. Also in keeping with the patterns seen within the reading cohorts, controlling for the RTM effect resulted in somewhat different interpretations about the relative effectiveness of the project among specific subgroups within the writing cohorts. Once again, groups that usually have higher baseline scores showed larger added value effects. For example, male students progressed at a nominally equivalent rate
to females within each cohort as indicated by the raw SDU, but after controlling for the RTM effect, the rate of progress for females is estimated to be approximately 30% greater than for males in both writing cohorts. In addition, the apparently higher rate of progress among Pasifika in the 2006-2007 cohort appears to have been due to the RTM effect, with the highest estimates of added value seen among NZ European students, and those of other ethnicities. However, the added value of the project remains fairly large among all subgroups, even after controlling for regression to the mean.

Table 16. LPDP added value within writing by subgroup after removing estimated RTM effect.

<table>
<thead>
<tr>
<th></th>
<th>Writing 2006-2007</th>
<th>Added value of the project</th>
<th>Writing 2008-2009</th>
<th>Added value of the project</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RTM effect (SDU)</td>
<td>Progress (SDU)</td>
<td>RTM effect (SDU)</td>
<td>Progress (SDU)</td>
</tr>
<tr>
<td>Overall</td>
<td>0.20</td>
<td>0.62</td>
<td>0.54</td>
<td>0.39</td>
</tr>
<tr>
<td>Male</td>
<td>0.31</td>
<td>0.65</td>
<td>0.46</td>
<td>0.48</td>
</tr>
<tr>
<td>Female</td>
<td>0.10</td>
<td>0.62</td>
<td>0.60</td>
<td>0.31</td>
</tr>
<tr>
<td>Low Decile</td>
<td>0.37</td>
<td>0.75</td>
<td>0.48</td>
<td>0.55</td>
</tr>
<tr>
<td>Mid Decile</td>
<td>0.14</td>
<td>0.43</td>
<td>0.39</td>
<td>0.40</td>
</tr>
<tr>
<td>High Decile</td>
<td>0.01</td>
<td>0.78</td>
<td>0.77</td>
<td>0.18</td>
</tr>
<tr>
<td>NZ European</td>
<td>0.15</td>
<td>0.62</td>
<td>0.58</td>
<td>0.30</td>
</tr>
<tr>
<td>NZ Māori</td>
<td>0.32</td>
<td>0.60</td>
<td>0.39</td>
<td>0.51</td>
</tr>
<tr>
<td>Pasifika</td>
<td>0.33</td>
<td>0.70</td>
<td>0.49</td>
<td>0.55</td>
</tr>
<tr>
<td>Other</td>
<td>0.12</td>
<td>0.69</td>
<td>0.66</td>
<td>0.38</td>
</tr>
</tbody>
</table>

**Colour Legend**
- **Red** 0 SDU
- **Orange** +0.4SDU
- **Green** +0.8SDU

### 4.2.2. Overall regression models

As noted previously, the RTM effect is quantified in the previous section using an extension of the ordinary least squares regression procedure discussed in section 4.1.3. Specifically, the EOY2 achievement is predicted using baseline score as a predictor, with the degree of scatter...
Chapter 4 - Single-level Regression

(reliability) used to indicate the likely RTM effect. This section builds on these basic models by investigating the residual effect of several other factors (after controlling for baseline score). As with all previous analyses, the main factors investigated were gender, ethnicity and decile group. However, since regression modelling is amenable to inclusion of multiple (relevant) predictors, the effect of several other factors was considered also. These factors were school size, school type (full primary, contributing or intermediate), and the proportion of minority students attending each school. Year level was also included since the distance from the asTTle norms differed across year levels. Various interactions were also considered to investigate moderating effects, such as whether there are differences in the effect of school size among students of different ethnicities, but no significant interactions were identified so these are not discussed further. Non-linear relationships were also considered but, in all cases, the linear estimator was clearly the best fit.

Gender, decile, ethnicity and school type were entered as n-1 dummy variables. That is, ethnicity for example can be regarded as a polytomous categorical variable with 4 levels (n = 4), requiring three binary dummy variables to be created. This process is outlined and justified in a number of publications, and avoids perfect multi-collinearity (e.g., Brambor, Clark, & Golder, 2006; Fox, 2008). The category coded as 0 in all of the dummies is considered to be the reference category, with results being interpreted in this context. The achievement and progress profile of those coded “other” is very similar to students who identified as NZ European, so the number of categories was reduced to three, with NZ European and students of “other” ethnicities regarded as the reference categories (within New Zealand, those in the “other” category are predominantly of Asian descent). Contributing schools were the reference category for school type, since this is the most common school type. Decile group could conceivably be considered either a polytomous categorical variable, or a scale variable. As the relationship between achievement/progress by decile group is non-linear, the analysis was better served using dummy variables (treating it as categorical), with low decile (1-3) treated as the reference category, mid (4-7) and high (deciles 8-10) entered as binary categorical variables. Year level was entered in the same way, with Year 4 treated as the reference category.
As with all previous analyses, achievement was measured using the adjusted asTTle score, centred on the asTTle normative mean, whereby the score indicates the distance from the relevant norm. This process does not affect the variance, so the published standard deviation of 100 remains the best indicator of population variance. School size was obtained from a publically available Ministry of Education affiliated website (www.ero.govt.nz) and ranges from 20 to 679 students. School size was centred to the average school size of each cohort to ensure that the intercept remained interpretable (though the intercept is of relatively little importance), but was unstandardized to retain maximum variance. The proportion of minority students was entered in a similar manner, with the proportion of minority students entered as a percentage, but centred to the average minority proportion for the specific cohort.

Stepwise regression and incremental sums of squares methodologies were used to investigate the predictive validity of individual factors and their higher order interactions. As already discussed, stepwise procedures are criticised as being atheoretical and analogous to data dredging due to the possibility of over-fitting of a model since these procedures are more sensitive to sampling error variations than the enter method. As a result, apparent relationships may only be valid within the specific data set, resulting in limited generalizability, though some argue that these issues can be overcome with large datasets using split-half or two studies (Copas, 1983; Rencher & Pun, 1980). In the current case, sampling error variations are mitigated by the large number of students. In all cases, the ratio of students to predictors is greater than 200 to one, which results in an estimated $R^2$ shrinkage of .01, meaning that the estimates of variance explained are likely to be accurate (Osborne, 2000). In addition, since results are compared across subsequent cohorts, generalizability is less likely to be problematic since findings are already based on generalisation across cohorts (Tabachnik & Fidell, 2001). Consequently, if results suggest that the findings are overly specific to each dataset, this would be strong evidence of the inadequacy of the stepwise methods.

Model adequacy has been assessed using the Gauss-Markov theorem, which states that the ordinary least squares estimator provides the best estimate of the coefficients if the mean of the residuals is zero and there is no remaining correlation between the residuals and the predicted
scores (Chipman, 2011). Multi-collinearity was assessed using the variance inflation factor (VIF), with VIF less than ten typically considered adequate (i.e., Hair, Anderson, Tatham, & Black, 1995; Kennedy, 1992; Marquardt, 1970; Neter, Wasserman, & Kutner, 1989).

4.2.2.1. Reading Focus – 2006-2007 Cohort

The regression model for students attending schools with a reading focus during the 2006-2007 cohort of LPDP is shown in Table 17, below. The adjusted $R^2$ was .59, and the mean residual was zero, with no correlation between the residual and predicted scores. Multi-collinearity does not appear to be a particular issue, with the VIF generally low. The higher VIFs seen for the Intermediate and Year 7 factors are unsurprising; these factors are strongly significant, so the marginally acceptable VIFs are not a major threat to conclusions made from the model. Variables are shown in the table in order of importance – that is, variables that explained the greatest additional portion of the variance in EOY2 scores (minus asTTle norms) were added sequentially by the stepwise procedure, and are shown in this order. All of the variables described in the previous section were found to have an effect on scores at EOY2, with the exception of the Pasifika dummy variable. That is, scores among Pasifika students at EOY2 were not significantly different than students in the reference category for ethnicity (i.e., NZ European and "other"), after accounting for the variance explained by the other variables, so the Pasifika variable is not shown in the table. Note that the exclusion of the dummy variable for Pasifika simply means that the reference category for ethnicity is all students who did not identify as NZ Māori.
Table 17. *Final stepwise regression model showing the effect of each factor on EOY2 score against expectation (n=4,163) among students in the 2006-2007 reading cohort.*

<table>
<thead>
<tr>
<th></th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
</tr>
<tr>
<td>(Constant)</td>
<td>-39.512</td>
<td>3.676</td>
<td>-10.750</td>
</tr>
<tr>
<td>Baseline score – norm</td>
<td>.731</td>
<td>.013</td>
<td>.610</td>
</tr>
<tr>
<td>High Decile</td>
<td>47.969</td>
<td>3.800</td>
<td>.233</td>
</tr>
<tr>
<td>Māori</td>
<td>-24.605</td>
<td>2.668</td>
<td>-.100</td>
</tr>
<tr>
<td>Year 7</td>
<td>53.530</td>
<td>5.595</td>
<td>.268</td>
</tr>
<tr>
<td>Year 6</td>
<td>35.898</td>
<td>4.092</td>
<td>.102</td>
</tr>
<tr>
<td>Mid Decile</td>
<td>21.175</td>
<td>3.121</td>
<td>.108</td>
</tr>
<tr>
<td>Centred School Size(^a)</td>
<td>.045</td>
<td>.008</td>
<td>.092</td>
</tr>
<tr>
<td>Centred Minority(^b)</td>
<td>.303</td>
<td>.071</td>
<td>.064</td>
</tr>
<tr>
<td>Year 5</td>
<td>10.458</td>
<td>2.676</td>
<td>.047</td>
</tr>
<tr>
<td>Female</td>
<td>7.388</td>
<td>1.966</td>
<td>.038</td>
</tr>
<tr>
<td>Full Primary</td>
<td>9.308</td>
<td>3.613</td>
<td>.038</td>
</tr>
</tbody>
</table>

\(^a\) Average school size in this cohort was 225 students
\(^b\) The average proportion of minority students was 35%

Even after controlling for baseline score, students in mid and high decile schools had higher achievement at EOY2 than those in the reference category (low decile), with an added SDU effect of approximately one-fifth and one-half a standard deviation, respectively. As with the regression to the mean section, this can be considered the “added value” of the project for students of these particular subgroups, and is likely to be a more accurate estimate since the model incorporates the variance explained by other factors. Students attending full primary and intermediate schools also typically had significantly higher achievement at EOY2 than those attending contributing schools, though this effect was of marginal practical significance for students attending full primary schools. Older students also typically had higher scores at EOY2.
Since the scores are centred by subtracting the expected asTTle score from each student’s actual score, and the baseline score has been incorporated within the model this suggests that older students gained more from the project.

Students who identified as Māori had lower EOY2 scores (SDU = 0.25) than students of other ethnicities, while students attending schools with a higher proportion of minority students had higher follow-up scores, though this effect was not large. Students attending schools with twice as many minority students as the average (i.e., schools where 70% of students are of NZ Māori or Pasifika descent) could be expected to have scores approximately ten points higher than students in the average school. School size was also of modest importance; students attending smaller schools (~100 students) typically scored around 20 points lower than those in large schools (~500 students). There were marginal gender differences, with female students gaining slightly more than males – however, while statistically significant, a difference of seven asTTle points is of little practical significance.

4.2.2.2. Reading Focus – 2008-2009 Cohort

As discussed in Chapter 2, this cohort differed somewhat from the other three clusters in that not all schools chose to use asTTle reading to measure achievement and progress. As a result, there were fewer students whose data could be included in the regression model, resulting in less coverage of variance within smaller subgroups. This meant that there were no intermediate schools represented in the data, so this could not be included as a factor, and no difference was found between contributing and full primary schools, meaning that school type had no effect in this cohort. Several other variables were also found to have a non-significant effect on achievement at EOY2, including gender, ethnicity, and school size (the average school size was 290 students in this cohort). The adjusted $R^2$ was a little lower, at .54, and the mean residual was zero, with no correlation between the residual and predicted scores, meeting the conditions of the Gauss-Markov theorem. The degree of multi-collinearity is low, with the VIF less than four for all
factors. Factors which predicted a significant portion of the variance in EOY2 scores (minus asTTle norms) are shown below in Table 18, in order of variance explained.

Table 18. Final stepwise regression model showing the effect of each factor on EOY2 score against expectation (n=1,510) among students in the 2008-2009 reading cohort.

<table>
<thead>
<tr>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
</tr>
<tr>
<td>(Constant)</td>
<td>-10.334</td>
<td>.4.273</td>
</tr>
<tr>
<td>Baseline score – norm</td>
<td>.657</td>
<td>.0.021</td>
</tr>
<tr>
<td>Year 7</td>
<td>102.88</td>
<td>5.348</td>
</tr>
<tr>
<td>Centred Minoritya</td>
<td>-.338</td>
<td>.110</td>
</tr>
<tr>
<td>Year 6</td>
<td>64.145</td>
<td>5.526</td>
</tr>
<tr>
<td>Year 5</td>
<td>17.368</td>
<td>4.234</td>
</tr>
<tr>
<td>Mid Decile</td>
<td>20.930</td>
<td>5.545</td>
</tr>
</tbody>
</table>

*The average proportion of minority students was 45%*

As with the previous reading cohort, older students typically had higher EOY2 scores – in the 2008-2009 cohort, these differences are large, with the added value of the project around one hundred points greater for students in Year 7 than for those in Year 4. There was also a significant decile effect, with students in mid and high decile schools typically gaining an additional 20 to 30 points by the EOY2 assessment, after controlling for baseline scores. Despite the finding that there were no significant differences in the value of the project for students of different ethnicities, there was a smaller gain among students attending schools with a higher proportion of students of NZ Māori and Pasifika descent. This result is the opposite (and of similar magnitude) to that found in the previous reading cohort, with a maximum difference of approximately 25 points (comparing a school where 10% of students are of minority ethnicity with a school where 90% are).
4.2.2.3. Writing Focus – 2006-2007 Cohort

The regression model for students attending schools with a writing focus during the 2006-2007 cohort of LPDP is shown in Table 19, below, in order of variance explained. There was no effect of school type for this cohort, but the same age-related pattern was apparent, with older students gaining more from the project. There was also no significant effect of being Pasifika, attending a mid-decile school, or among schools of different sizes (the average school size was 260). The non-significant factors are not included within the regression model, meaning that the reference groups of the dummy variables relating to ethnicity and decile represent all students not explicitly included in the model. Specifically, the reference group for NZ Māori is all non-Māori, and the comparison group for students attending high decile schools is made up of students in all other deciles (1 to 7). The adjusted $R^2$ was lower than for the reading clusters at .4. The mean residual was zero, with no correlation between the residual and predicted scores, meeting the conditions of the Gauss-Markov theorem. The degree of multi-collinearity is low, with the VIF close to one (which is the minimum) for all factors.
Table 19. Final stepwise regression model showing the effect of each factor on EOY2 score against expectation (n=3,700) among students in the 2006-2007 writing cohort.

<table>
<thead>
<tr>
<th></th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
</tr>
<tr>
<td>(Constant)</td>
<td>17.567</td>
<td>3.169</td>
<td></td>
</tr>
<tr>
<td>Baseline score – norm</td>
<td>.438</td>
<td>.012</td>
<td>.507</td>
</tr>
<tr>
<td>High Decile</td>
<td>33.730</td>
<td>3.865</td>
<td>.139</td>
</tr>
<tr>
<td>Year 7</td>
<td>46.180</td>
<td>4.576</td>
<td>.142</td>
</tr>
<tr>
<td>Female</td>
<td>22.206</td>
<td>2.860</td>
<td>.102</td>
</tr>
<tr>
<td>Māori</td>
<td>-15.366</td>
<td>3.536</td>
<td>-.060</td>
</tr>
<tr>
<td>Year 6</td>
<td>27.158</td>
<td>4.896</td>
<td>.077</td>
</tr>
<tr>
<td>Year 5</td>
<td>7.958</td>
<td>3.207</td>
<td>.035</td>
</tr>
<tr>
<td>Centred Minoritya</td>
<td>-1.148</td>
<td>.070</td>
<td>-.035</td>
</tr>
</tbody>
</table>

*a The average proportion of minority students was 35%

As noted above, older students typically had higher EOY2 scores after accounting for the variance explained by the other factors. The difference between Year 4 (the reference category) and Year 5 was not large, but students in Years 6 and 7 gained considerably more from the project (approximate SDU = .25 and .45, respectively). There was no significant difference between low and mid decile schools, but students in high decile schools gained approximately one-third of a standard deviation more than students in other deciles. Māori gained marginally less from the project than students of other ethnicities, and there was also a smaller gain among students attending schools with a higher proportion of students of NZ Māori and Pasifika descent, though the effect is very small in the current cohort – approximately ten points across the range of schools. Female students also typically gained more (approximate SDU = 0.2) from the project than male students.
4.2.2.4. Writing Focus – 2008-2009 Cohort

The regression model for students attending schools with a writing focus during the 2008-2009 cohort of LPDP is shown in Table 20, below, in order of variance explained. As with the previous writing cohort, there was no effect of school type. There was also no significant difference among students attending schools of different sizes (the average school size was 215). The non-significant factors are not included within the regression model, meaning that the reference groups of the dummy variables relating to decile and year level represent all students not explicitly included in the model. Specifically, the reference group for students attending mid decile schools is made up of students in all other deciles (i.e., those in low and high decile groups). The adjusted $R^2$ was .38, and the mean residual was zero, with no correlation between the residual and predicted scores, meaning that the OLS estimator is the most appropriate estimate of the coefficients according to the Gauss-Markov theorem. The degree of multicollinearity is low, with the VIF less than two for all factors.
Table 20. Final stepwise regression model showing the effect of each factor on EOY2 score against expectation (n=3,043) among students in the 2008-2009 writing cohort.

<table>
<thead>
<tr>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
</tr>
<tr>
<td>(Constant)</td>
<td>35.316</td>
<td>3.450</td>
</tr>
<tr>
<td>Baseline score – norm</td>
<td>.472</td>
<td>.014</td>
</tr>
<tr>
<td>Centred Minority*</td>
<td>-.190</td>
<td>.081</td>
</tr>
<tr>
<td>Female</td>
<td>20.742</td>
<td>3.196</td>
</tr>
<tr>
<td>Year 7</td>
<td>23.112</td>
<td>4.220</td>
</tr>
<tr>
<td>Māori</td>
<td>-18.799</td>
<td>4.382</td>
</tr>
<tr>
<td>Mid Decile</td>
<td>10.709</td>
<td>3.596</td>
</tr>
<tr>
<td>Year 6</td>
<td>13.375</td>
<td>5.013</td>
</tr>
<tr>
<td>Pasifika</td>
<td>-13.231</td>
<td>5.432</td>
</tr>
</tbody>
</table>

\* The average proportion of minority students was 38%

There was no significant difference between students in Years 4 and 5, while students in Years 6 and 7 gained more from the project than the younger students, but the differences were more modest than for any of the three other clusters of students. In this cohort, there was no significant difference between low and high decile schools, while students in mid decile schools gained marginally more than students in other deciles. This difference is very small (SDU = .1), so it would be reasonable to assert that students of all decile groups within this cohort gained a similar amount from the project. Māori and Pasifika gained marginally less from the project than students of other ethnicities, and there was also a smaller gain among students attending schools with a higher proportion of students of NZ Māori and Pasifika descent. However, as with the previous writing cohort, these effects are small. As with the previous writing cohort, female students typically gained more (approximate SDU = 0.2) from the project than male students.
4.2.3. Value-added models

This section uses the regression models developed in the previous section to explore whether there were differences in the degree of value-added by the project according to which facilitator was working with the set of schools (see section 2.1.1 for a description of the facilitation structure). This was assessed by comparing the average residual regression scores of students depending on the facilitator who worked with their school. This procedure is frequently used to assess the extent to which a school has “added value” relative to other schools with similar characteristics since systematic differences in residual scores are argued to show the effect of being in a particular group, after having taken any variables in the model into account (Jakubowski, 2008; Ray, Evans & McCormack, 2008). This means that differences in student progress by facilitator can be compared, with the understanding that any differences are not simply a result of the factors included in the models in subsection 4.2.2; that is, factors the facilitator cannot control, such as the students’ baseline scores, the decile of the schools the facilitator was working with, and so on. Therefore, if students whose schools were working with a particular facilitator tend to have residual scores that differ markedly from zero, one can assume that the facilitator involved contributed more or less (depending on whether the residual is positive or negative) than other facilitators.

The analysis relies on the premise that if there are systematic differences remaining within the residual scores, these differences are reflective of a real effect (Jakubowski, 2008; Ray et al., 2008). It is likely that there are other predictors that could be included in the regression models that would explain additional variance, but unless these mirror the facilitators in some way it is unlikely that this would explain any systematic differences. The difference between residual scores by facilitator was analysed using one-way ANOVAs of the residual scores by facilitator, with Tukey’s B post-hoc analyses to identify significant differences among facilitators. In order to be as confident as possible that any differences related to the facilitator rather than the school, the residual scores of students from both cohorts were combined and only facilitators with data for more than 100 students and more than three schools were included in the analysis. The regression models were comparable across time (i.e., 2006-2007 or 2008-2009), but differed
somewhat by focus (i.e., reading versus writing), so the analysis remained separated by focus. It was also hypothesised that certain facilitators might be stronger within one focus than the other, so the separate analysis also allowed this to be investigated.

4.2.3.1. Reading Focus – both cohorts

Among schools focusing on reading, there were 17 facilitators with data from more than three schools and at least 100 students, four of whom elicited gains significantly greater than the gains within the project overall (p < .05). Conversely, there were seven facilitators where students attending schools working with these facilitators achieved significantly lower EOY2 scores than predicted after accounting for the factors included in the regression model (p < .05), while the remaining six achieved progress equivalent to the project overall. The differences were moderate, with students attending schools working with the four facilitators identified as producing larger gains progressing 0.4SDU more than students in schools working with the seven facilitators whose students progressed more slowly than predicted.

However, for three of the seven facilitators working with schools whose students made lower gains, the actual average discrepancy, although statistically significant, was very small (5–8 points). In order to investigate whether these differences were consistent, the school-level results are also considered. Schools working with these three facilitators did not have consistently lower effects (see Table 21), indicating that the overall lower progress for students whose schools were working with these facilitators was unlikely to be educationally significant. However, the gains tended to be consistently lower for schools working with the remaining four facilitators whose students had larger differences between predicted and actual results (16–25 points), with two of these four showing significantly lower than predicted gains for all of the schools working with these facilitators. Students attending schools working with the remaining two facilitators in this group (whose average difference from predicted was 16-19 points) had significantly lower than predicted gains for more than half of their schools.
Table 21. Consistency of the shifts among the reading schools working with facilitators with lower added value.

<table>
<thead>
<tr>
<th>Average negative shift across schools</th>
<th>N</th>
<th>Consistency across schools working with each facilitator</th>
</tr>
</thead>
<tbody>
<tr>
<td>5–8 points</td>
<td>3</td>
<td>Not consistently lower</td>
</tr>
<tr>
<td>17-25 points</td>
<td>2</td>
<td>All schools consistently lower</td>
</tr>
<tr>
<td>19 points</td>
<td>1</td>
<td>3 schools significantly lower</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 school average</td>
</tr>
<tr>
<td>16 points</td>
<td>1</td>
<td>3 schools significantly lower</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 school average</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 school significantly higher</td>
</tr>
</tbody>
</table>

One of the two facilitators working with reading schools where students progressed more than the project as a whole worked with two schools where students gained more than predicted from the project, and one school in which students gained about the same as students in the project overall. The other facilitator, whose schools had the largest additional average gain over prediction (25 points), worked with seven schools in total. In four of these schools, students made gains significantly greater than predicted; in two schools, students made gains similar to the overall project; and in one, students’ gains were lower than the average for the project. This is described in Table 22.
Table 22. *Consistency of the shifts among the reading schools working with facilitators with higher added value.*

<table>
<thead>
<tr>
<th>Average positive shift across schools</th>
<th>N</th>
<th>Consistency across schools working with each facilitator</th>
</tr>
</thead>
<tbody>
<tr>
<td>18 points</td>
<td>1</td>
<td>2 schools significantly higher</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 school average</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 schools significantly higher</td>
</tr>
<tr>
<td>25 points</td>
<td>1</td>
<td>2 schools average</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 school significantly lower</td>
</tr>
</tbody>
</table>

4.2.3.2. Writing Focus – both cohorts

Among schools focusing on writing, there were 18 facilitators with data from more than three schools and at least 100 students, five of whom elicited gains significantly greater than the gains within the project overall ($p < .05$). Conversely, there were three facilitators where students attending schools working with these facilitators achieved significantly lower EOY2 scores than predicted after accounting for the factors included in the regression model ($p < .05$), while the remaining ten achieved progress equivalent to the project overall. The differences were again moderate, with students attending schools working with the five facilitators identified as producing larger gains making progress almost half a standard deviation greater than students in schools working with the three facilitators whose students progressed more slowly than predicted.

In writing schools, the facilitator working with schools where students made the lowest gains compared to prediction (on average, 43 points less progress than the overall project) had results that were consistent across these schools (see Table 23). This facilitator worked with eight schools over the two cohorts. Seven of these schools had lower progress than predicted though this result was only statistically significant in four schools (the three schools where this was non-significant were small, so are collectively considered evidence of a trend). Students attending the remaining school that worked with this facilitator had progress no different than predicted. The remaining two facilitators, working in schools where students showed lower gains, had more mixed results. The facilitator with the second largest difference between predicted and actual progress (22 points lower) worked with five schools: two of these progressed significantly less
than predicted, two had progress no different to the overall project, and one had progress significantly higher than predicted. Students working with the remaining facilitator gained an average of 13 points less than predicted. The three schools working with this facilitator had average gains slightly lower than the project overall, but these differences were non-significant for each school (though significant overall).

Table 23. Consistency of the shifts among the writing schools working with facilitators with lower added value.

<table>
<thead>
<tr>
<th>Average negative shift across schools</th>
<th>N</th>
<th>Consistency across the schools each facilitator worked with</th>
</tr>
</thead>
<tbody>
<tr>
<td>43 points</td>
<td>1</td>
<td>4 schools significantly lower</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 small schools lower</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 school not significantly different from average</td>
</tr>
<tr>
<td>13–22 points</td>
<td>2</td>
<td>Inconsistent</td>
</tr>
</tbody>
</table>

There were five facilitators who worked with schools that made higher than predicted gains in writing (see Table 24). The first facilitator (average difference of 18 points) worked with nine schools, but progress among these schools was variable. Students attending four of these schools progressed significantly more than predicted, three made gains that did not significantly differ from the project overall (though two were marginally higher than predicted), while students attending the remaining two made significantly less than predicted progress. One of these schools actually had the largest difference between actual and predicted progress in the writing sample, with students making average progress 84 points below what had been predicted based on the factors entered into the regression models. Given that this school was markedly different from all other schools it is likely that at least some of this effect was related to school-level factors that were not captured by the regression models (e.g., a particularly ineffective principal, or changes in the testing procedures).

The results for the remaining four facilitators working with schools where students typically made greater than predicted gains in writing (21–23 points) tended to be consistent. Students
working with one of these four facilitators had average gains 21 points more than predicted by the regression models. All four of the schools working with this facilitator showed higher than predicted progress, though this was not statistically significant in one school due to the low number of students. The second facilitator (difference of 21 points also) had higher than predicted gains in six out of seven schools, though for two of these schools the difference was non-significant (again, these schools were small), while the seventh school working with this facilitator had progress similar to that predicted by the project overall. The three schools working with the third facilitator (again with a difference of 21 points) all progressed more than predicted, though this was not significant in two of the schools since they were relatively small schools. For the fourth facilitator (average difference 23), progress was higher than predicted in three of the six schools, and not significantly different in the other three.

Table 24. Consistency of the shifts among the writing schools working with facilitators with higher added value.

<table>
<thead>
<tr>
<th>Average positive shift across schools</th>
<th>Number of facilitators</th>
<th>Consistency across the schools each facilitator worked with</th>
</tr>
</thead>
<tbody>
<tr>
<td>18 points</td>
<td>1</td>
<td>Inconsistent, 9 schools</td>
</tr>
<tr>
<td>21 points</td>
<td>2</td>
<td>8 schools significantly higher</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 schools higher (not significantly due to size)</td>
</tr>
<tr>
<td>21 points</td>
<td>1</td>
<td>4 schools significantly higher</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 schools higher</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 school average</td>
</tr>
<tr>
<td>23 points</td>
<td>1</td>
<td>3 schools significantly higher</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 schools average</td>
</tr>
</tbody>
</table>

There were some unusual patterns that surfaced in the results of the facilitation effect. For example, of the four facilitators for whom student progress was better than average in writing, only one facilitator elicited better than average gains among students in reading. Conversely, no facilitator was in the worse than average group for both foci, while one facilitator achieved better than average progress for writing students, but worse than average for those in reading schools.
In addition, the presence of several instances in which the shifts differed across the facilitators’ schools, indicates that there is likely to be residual inter-school variation that is outside the facilitator’s realm of influence. This variation could be due to a number of single or multiple factors, such as differences in the extent to which principals and teachers in a particular school were amenable to changing practice based on the professional development provided by facilitators. However, for facilitators whose overall average positive difference between predicted and actual progress is especially large, there is distinctly higher inter-school consistency, implying a facilitator effect unique to these individuals.

4.3. Chapter Summary

The use of regression models allows considerable contextual information to be incorporated, supporting a more nuanced exploration of the factors that relate to different levels of achievement and progress among different groups of students, compared with the pairwise effect size comparisons in Chapter 3. The specific analyses focusing on regression to the mean indicate that conclusions about the relative value of the project differed, sometimes markedly, compared with the results of the effect size analyses which are procedures that do not take RTM into account. In contrast, conclusions about the relative gain of subgroups derived from the more general regression models were generally similar to the RTM analyses. However, despite the overall similarities, the more general models provide a more complete picture of the effect of being in a particular subgroup within the project, since the effects are estimated with respect to the variance explained by all of the other variables included in the model. Indeed, the similarities between the results of the RTM and general regression models arise because the general models implicitly account for the RTM effect by incorporating the baseline score as a covariate – and baseline score is the strongest predictor of achievement at EOY2. This relationship is frequently misinterpreted within the literature, with a slope less than one for the baseline score factor taken as evidence that students with the lowest achievement progressed more than students with higher achievement (Lohman & Korb, 2006; Smith & Smith, 2005). While this is indeed what the
model suggests, the reasons for this are frequently statistical rather than educationally meaningful, since the apparent “regression” of scores may be wholly explained by the RTM effect. It is difficult to partial out the relative impact of the intervention compared with the expected regression due to RTM once multiple factors are included in the model, so the specific RTM analyses are considered the best quantification of the average project value, while the regression models provide more information about the relative value among subgroups.

The validity of the RTM analyses is reliant on the extent to which the asTTle normative mean is reflective of the true population mean for the students within the LPDP cohorts. Within the writing sample in particular, it is likely that the asTTle norms overestimate the true population mean for these cohorts of students, since schools opted into the project based on their own perceived need for additional professional learning rather than being selected due to low test results. As a result, the true population mean for these students is highly likely to be lower than the asTTle norms so the RTM results are likely to be a very conservative estimate of the added value of the project, particularly in writing. Despite this conservative procedure, the project value, after accounting for the RTM effect, was estimated to improve from .34 to .39 SDU in reading, and from .54 to .61 SDU in writing. Hattie (2009) suggests that a mean change in SDUs equal to 0.4 is sufficiently large to warrant educational significance, but the RTM effect was not considered within Hattie’s synthesis of meta-analyses so these effects are likely to be even more significant. These results are in contrast with the effect size analyses presented in Chapter 3, in which estimated effects were more than twice as large for writing as for reading. Nevertheless, the results suggest an improvement for subsequent cohorts, with the latter cohort (2008-2009) achieving greater gains than those in the 2006-2007 cohort, and this improvement is consistently suggested by the results of all analyses presented thus far.

Although the results of the general regression models suggest some differences in the gains among schools of different types, sizes and proportion of minority students, the actual magnitude of the effects is relatively small and unlikely to be meaningful for educational policy, especially since it does not account for the numerous other advantages (and possible disadvantages) of attending schools of different size and type. Of concern, however, were the findings regarding the
influence of socio-economic factors and of ethnicity; while the intervention appears to have facilitated good progress overall, these analyses show it to be generally more successful for groups that are traditionally advantaged. This is obviously concerning since it suggests that even in a successful intervention where under-achievers are especially targeted (and successful), those in minority and otherwise disadvantaged groups still make lesser gains than other students with similar initial achievement.

The value-added analyses indicated a possible facilitator effect, with some facilitators adding more or less value, when compared with the average progress for the entire cohort. It is important to consider these results within the context of the project overall. LPDP appears to have been a very successful project for the majority of students involved and much of the credit for this success is directly related to the skill and dedication of the facilitators, who worked directly with the schools. Therefore, the results of this analysis should be seen within the context of a successful intervention – that is, the majority of facilitators achieved good results, but some were able to bring more to the project than others. Typically this effect was modest, but in some cases was large enough that further investigation would be warranted if the project had been ongoing. In reading, the SDU between the facilitators with the smallest added value compared with those with the most was approximately half a standard deviation, while for writing there was a difference of up to three-quarters of a standard deviation. Interestingly, only one facilitator achieved better than average progress for both reading and writing. No facilitator was in the worse than average group for both foci, and one facilitator achieved better than average progress for writing students, but worse than average for those in reading schools. This may indicate that facilitators tend to achieve better results in one focus than the other; potentially implying that facilitator specialisation could be useful for any future interventions. However, the validity of the value-added analyses is reliant on the extent to which possible covariance has been captured. As noted in section 4.2.3.2, one facilitator who worked with nine schools had apparently added significantly more value than the project overall, but this effect was inconsistent. Further investigation revealed that the mean residual for one of the nine schools was vastly greater than all other schools, raising the
likelihood that there was unique variance related to the school itself, rather than the facilitator, that had not been captured.
CHAPTER 5. HIERARCHICAL LINEAR MODELS

The previous chapter reviewed the concepts of single level regression related to the analyses conducted on the LPDP data. Basic models investigating the effect of regression to the mean were considered, followed by the development of models incorporating the various demographic covariates collected during the project. The final models were then used to investigate differential effectiveness among individual facilitators. The results indicated some substantive differences compared to the effect size analyses, though the different regression procedures yielded comparatively similar results. The current chapter presents the final set of analyses of the LPDP data, using hierarchically structured regression analysis, whereby partial dependency in the data is appropriately accounted for in the models. The results of the analyses presented in this chapter can be easily compared to the analyses in the previous chapters, since these too, can be represented as standard deviation units (SDU). The overall models presented in subsection 4.2.2, as well as the results of the value added models presented in the subsequent subsection, are especially comparable to the analyses in the current chapter, since these differ only in that the current models incorporate the hierarchical structure of the data. In addition, two different estimation procedures are presented, to determine the extent to which these differ. These comparisons address the primary aim of the thesis, while the secondary intention to investigate the effectiveness of the project requires the results to be considered more holistically.

The structure of Chapter 5 is similar to the other analysis chapters, whereby the first section of the chapter provides a general review of the hierarchical linear modelling (HLM) concepts (section 5.1) that underpin the analyses conducted in the analysis section (section 5.2). The first subsection of this review introduces the concept of hierarchically structured data (subsection 5.1.1), which is followed by an explanation of why this structure should be incorporated in model development (5.1.2). Subsection 5.1.3 discusses the origins of the equations required to conduct HLM analysis, while subsection 5.1.4 illustrates how to calculate the degree of dependency in the data. This calculation is of particular importance since this provides the main motivation for conducting hierarchical analysis. The steps that were presented in the previous chapter
Hierarchical linear models (subsection 4.1.3) are applied to the multilevel framework in subsection 5.1.5 to emphasise the analogousness of the two methods, and to support the procedures used in the analyses presented in the chapter. Due to the relative complexity of HLM, there were additional considerations that needed to be introduced in subsection 5.1.6, especially with respect to longitudinal analysis and centring procedures, both of which were important considerations in the current context.

The second major section of the chapter is the analysis section. Since all subsections of the analysis use the same overall procedures and models (though with different data), details about the model development are provided in the general analysis discussion (section 5.2). Each set of analyses is then presented in a separate subsection, with the overall models for each of the four independent datasets (reading 2006-2007; 2008-2009; writing 2006-2007; 2008-2009) presented in subsections 5.2.1 through 5.2.4. Subsection 5.2.5 presents the results of the value-added models based on the overall models developed in the previous subsections. The final section of the chapter (5.3) summarises the results and aggregates the findings with particular attention to any differences between the methods. Again this focuses primarily on differences that arise within the chapter, with the main discussion of differences between chapters presented in Chapter 6, but findings of particular significance are indicated.

5.1. Introduction to the HLM concepts used in this chapter

This section introduces to the basic concepts and statistics used in HLM analysis. Within the constraints of the thesis structure this cannot be an exhaustive introduction, but the concepts required to understand the analyses developed within this chapter are explained, hence the focus specifically on HLM rather than multilevel models in general. For a more complete introduction, see “Data Analysis using Regression and Multilevel / Hierarchical Models” by Gelman and Hill (2007), or “A User’s Guide to MLwiN” by Rasbash, Steele, Browne and Goldstein (2012) both of which were the main sources used within this section.

---

9 HLM is a subset of multilevel modelling, and is the multilevel analogue of linear regression within the single-level regression framework.
5.1.1. Hierarchically structured data

Hierarchical linear models (HLMs) are an extension of single-level linear regression models such as those described in Chapter 4, and just as in conventional regression, HLM estimates the effect of one or more predictors on the outcome variable. However, HLM differs in that the hierarchical model incorporates information about the underlying data structure in its estimates, allowing the degree of variance at each level of the hierarchy to be assessed, and retaining the hierarchical structure when estimating the effect of predictors for which data are available at different levels of the hierarchy. This is convenient, in that it allows the analysis to reflect a more naturalistic framework since data frequently reflect a hierarchical structure in the social sciences, especially within the educational context (Osborne, 2000). However, the primary argument for doing so is that there is likely to be a degree of similarity (or dependence) among students attending a particular school that is not shared or explained across contexts. Therefore, within the hierarchical framework, students can be conceptualised as nested within schools, reflecting two levels of hierarchy, with students as level one, and schools as level two. Students in the same school are likely to have a range of shared experiences (e.g., same teacher, school rules, shared homework, and discussions at lunch) that may partially explain differences in outcomes across students attending different schools.

The two-level hierarchy described above is not the only possible representation however. If measurements are repeated over time, individual observations or tests relating to each student nested within that student could be considered level-1 units, making students themselves level two and schools level three (Rasbash et al., 2012). Individual classrooms could also be included as an additional level, as could region or cluster (schools in NZ often work closely with other ‘similar’ schools). Typically the researcher conceptualises the hierarchical structure based on theoretical and practical knowledge (Griffith, 2002), and then tests this structure during the model building stage. The degree of clustering/dependency is estimable using the Intra Class Correlation (ICC), and provides an indication of the degree of risk inherent in excluding the relevant level from the model. The ICC is discussed further in Section 5.1.4. There are also procedures available that allow the hierarchical structure (number and order of levels) to be data-
determined prior to running an HLM (e.g., Within and Between Analysis, or WABA), but such methods are most useful when the hierarchical structure of the data is unclear, which is typically not an issue in the educational case (Dansereau, Cho & Yammarino, 2006).

5.1.2. Reasons for incorporating hierarchical structure

At a simple level, the inclusion of the hierarchical structure in the regression equation is necessary since there is a degree of shared covariance among individuals within each level of the hierarchy; students attending the same school are somewhat more likely to share certain similarities than students attending different schools, such as more similar backgrounds, experiences and teaching (Clarke, 2008; Gelman & Hill, 2007; Seltzer, & Rickles, 2012; Woltman, Feldstain, MacKay, & Rocchi, 2012). HLM allows the researcher to investigate the effect of higher-level predictors on the outcome variable efficiently.

In conventional regression, researchers often encounter and succumb to the “fallacy of the wrong level”, described cogently by Dansereau and colleagues (2006). This fallacy refers to the use of results from a certain level (e.g., students) to make inferences about the influence of predictors at another level (e.g., school). For example, when attempting to determine which teachers/schools are the most effective, researchers often aggregate individual-level data to derive a school-level variable and attempt to investigate shifts at the school-level (Griffith, 2002). The rationale for using such aggregates is questioned however, since such aggregates assume homogeneity in the data that comprise each aggregate (which rarely exists) and elide all of the variance in the individual-level data (Griffith, 2002). Value-added models (such as those presented in Chapter 4) typically retain the individual-level variability, but are also regarded as problematic since the estimates of effectiveness are derived from mean aggregates of the individual-level residuals – that is, school-level effectiveness is still being determined by student-level data (Dansereau et al., 2006; Griffith, 2002). The fundamental issue in using individual-level data (e.g., achievement data) to determine school-level attributes (e.g., effectiveness), relates to the extent to which the individual-level data reflect school-level characteristics. There is frequently
no empirical/theoretical justification for assuming that this would be the case (Griffith, 2002), and
failure to account for the possibility that the relationships between variables may differ across
levels can result in the atomistic, or ecological fallacy (Diez-Roux, 1998; 2002).

The atomistic fallacy occurs when conclusions are made about variability across groups
based on individual-level data, while the ecological fallacy occurs when making conclusions
about individuals, based on group-level data. In its extreme, the fallacy of the wrong level can
result in conclusions that are the polar opposite of what the data actually describe (Gelman & Hill,
2007). This becomes increasingly likely as the number of groups (e.g., schools), or group-level
variability increases. A real-world and intuitive example of the possible ramifications of the fallacy
of the wrong level comes from the political arena. In the United States, the wealthiest states are
more likely to vote Democrat. If one were to extrapolate this, and conclude that wealthy
individuals must also be more likely to vote Democrat, this would be an instance of the ecological
fallacy resulting in the opposite conclusion, since, at the individual level, wealthier people are
actually more likely to vote Republican (Gelman, Park, Shor, Bafumi & Cortina, 2008).

Using the above example, it should be evident that it would be possible to reach the correct
conclusions using conventional regression if one is careful about the model specification and the
level at which conclusions are being made, either by running separate regressions for each state,
or by aggregating individuals within each state. However, this process can become rapidly very
complex when additional predictors are included, especially when these predictors are at different
levels – a common situation within schooling research. For example, in the LPDP sample, SES is
a school-level predictor, while ethnicity is an individual-level factor. A common research question
might be to determine whether achievement and / or progress are affected by an individual’s
ethnicity after taking school-level SES into account. The use of conventional regression to
investigate this question would tend to inappropriately reduce the standard errors of parameter
estimates leading to increased risk of type-I error, and may obfuscate possible differences in the
nature of the relationship between the predictors and the outcome variable across different
contexts (in this case, schools) (Seltzer & Rickles, 2012).
Conversely, the use of HLM affords a number of advantages since the complete set of analyses are built into the overall model concurrently, thereby accounting for the shared variance inherent in multilevel data, and allowing conclusions to be made about the parameters at each level in the hierarchical structure. As with all quantitative methodologies, there are certain statistical assumptions inherent in HLM which diminish its relative advantage as these assumptions become more seriously violated. In such instances, single-level regression can be used with little risk to the validity of our conclusions. However, the statistical assumptions of HLM are inherently the same as for conventional regression; that is, of linearity, independence, equal variance and normality, except with HLM, the assumptions apply at each level of the analysis (Gelman & Hill, 2007). As a result, some authors assert that the hierarchical structure is so integral to accurate analysis and interpretation of such data, that multilevel modelling techniques such as HLM should invariably be used over single-level regression techniques (e.g., Clarke, 2008; Gelman & Hill, 2007). Indeed, using simulation studies, Clarke (2008) demonstrates that ignoring the clustering within the data leads to an increased risk of type I error even when the clustering is slight, and the data at each level are sparse (as few as two observations per group). Further, it is argued that analysing the data using HLM will, at worst, result in a model with results no more accurate than those generated using conventional regression (Clarke, 2008; Gelman & Hill, 2007). This provides a very strong argument for multilevel analysis wherever practical.

Despite the strong endorsement of multilevel techniques in the methodological literature, conventional regression remains much more popular than all forms of multilevel modelling combined, despite an increase in the availability of software to perform such analyses (SPSS, Mplus, R, SAS and MLwiN all allow multilevel structures to be analysed, though with varying degrees of functionality and limitations). A search of major databases (PsycInfo and ERIC) comparing the number of hits using the keywords “regression model”, versus the sum of hits for the keywords “multilevel model” and “hierarchical linear model”, indicates that regression models are still approximately eight times more common than HLM and multilevel modelling collectively. Obviously there are a number of issues with such an estimate, in that the terms used to describe

---

10 This search was limited to results from 2005 to 2013
multilevel modelling techniques are much more variable. For example, some authors use terms such as “multilevel regression model” which would result in a small degree of inflation in the number of articles that appear to fall into the conventional regression category. However, these articles would appear in both categories – meaning that even if absolutely every multilevel article used this term, there would still be seven times more conventional regression articles. There are a number of other issues that would affect the precision of this estimate, but given the sample size (more than 30,000 articles) and the overwhelming difference, it is reasonable to conclude that conventional regression remains considerably more common than HLM. This relative difference in usage is likely to be primarily due to the increased complexity of HLM, which requires a level of statistical knowledge beyond the capabilities of most researchers and graduate students in the social sciences, at least in the English speaking world (Byrne, 2012; Clay, 2005; Cotterell & von Randow, 2010; MacInnes, 2009; Wiles, Durrant, De Broe, & Powell, 2009). Added to this is the perception that HLM requires a very large dataset, which is often not realistic given practical and financial restrictions on the research; though as previously indicated this perception is unwarranted (Clarke, 2008).

5.1.3. The origins of HLM

As previously noted, HLM is simply a more complex representation of the conventional, ordinary least squares regression discussed in Chapter 4, except that HLM retains and incorporates the hierarchical structure and shared variance at each level into the analysis. The first algorithm demonstrated to allow estimation of covariance components from hierarchical data was presented more than thirty years ago (Woltman et al., 2012), and is commonly known as the Expectation-Maximisation (EM) algorithm, first presented by Dempster, Laird and Rubin (1977). EM is used extensively for a variety of applications, but is most well recognised in the educational context for its use in providing an estimate for missing data values, which was also the original application presented by Dempster and colleagues. However, estimation of a covariance parameter is analogous to deriving an estimate for a missing datum and, in 1981, it was
demonstrated that EM could be used to produce parameter estimates within the context of hierarchical linear regression (Dempster, Rubin, & Tsutakawa, 1981), paving the way for efficient calculation of HLMs (Woltman et al., 2012).

5.1.4. Assessing the dependency assumption

A fundamental assumption of HLM is that there is a degree of dependency in the data, which would violate the assumption made by conventional regression analysis that each individual within the analysis is independent. For example, it is highly likely that the achievement of students attending a particular school will be somewhat dependent on the group/cluster they belong to (in this case, the school), as opposed to a different cluster. To quantify the degree of dependence explained by the clustering in the data (e.g., classrooms, schools, regional clusters), the Intraclass Correlation (ICC) is used. The ICC reflects the proportion of the total variance that is explained by group. It is defined as:

\[
 ICC = \frac{\tau}{\tau + \sigma^2}
\]

(11)

where: \( \tau \) = between group variance; \( \sigma^2 \) = pooled within group variance; \( \tau + \sigma^2 \) = total variance in the outcome variable, \( Y \).

To illustrate the use of the ICC, the example described in Section 4.1.2 is presented. To remind the reader, in this example, a random 10% sample of Year 7 students was selected, and the relationship between their baseline and follow-up achievement investigated using a basic regression model, ignoring any possible nesting in the data. For simplicity, time two achievement is the outcome variable as it was in the single level regression example, considering only those students for whom both Time 1 and Time 2 data were available (\( n = 190 \)), and only two levels described here (individuals within schools). There are several ways to estimate the between and within group variances required for the ICC calculation but, in this case, an unconditional model (one without any predictors) was specified using the MLwiN software package, using the Restricted Iterative Generalized Least Squares (RIGLS) estimator. RIGLS is analogous to
Restricted Maximum Likelihood, and has been demonstrated to be an unbiased estimator (Goldstein, 1989). The between group variance was estimated as 1319.9, and the pooled within group variance as 5479.5, yielding an ICC of 19%. This suggests that almost one-fifth of the variability in scores at Time 2 is at the school-level, which represents considerable clustering in the data. As indicated earlier, it is possible to conduct HLM even if the dependency in the data is very low – indeed, if the ICC is 0, the results will be the same as for single level regression. However, as the degree of dependency increases, the threat to the validity of conclusions made using conventional regression increases rapidly. With a relatively low ICC of .1, and a small study of 90 participants across three equal groups, the alpha (likelihood of an apparently significant result due to chance) would be .492 rather than the conventional .05 (SAS, n.d.).

The ICC is a useful guide in determining how many levels to include in the model, but as the number of levels increases, so too does the complexity of the HLM. To demonstrate this, the following theoretical figures indicate the relationships that could be estimated at each level of the simple two-level example. For ease of interpretation, these figures can be considered to represent students nested within schools. Figure 7 shows that at the student level (Level 1) the relationship being estimated looks no different than for conventional regression. However, in HLM student level variation is estimated with respect to the school in which the students are situated, so Figure 7 represents students in a single school, and each school would have its own regression line fitted, possibly with different slopes and / or intercepts.
Figure 7. Level 1 regression line for an individual school.

Figure 8 shows a case where the intercepts of individual schools are allowed to vary, but where individual school regression slopes are fixed. When the school-level regression lines are fixed for each school, the intercept of each school’s regression line varies, indicating that students with the same baseline scores would be expected to gain more in some schools than others, and this “added-value” effect would be equivalent across the distribution of T1 scores; allowing no random variation at the school-level. This is useful when the researcher is primarily interested in the student-level relationships and wishes to retain simplicity in the model.
Figure 8. Level 2 regression lines for individual schools with varying intercepts but fixed slopes.

Allowing the slopes to vary incorporates more information about the underlying data structure, and is generally advantageous to explore this since it is frequently a more realistic representation of the data. In addition, it allows the researcher to investigate school-level variation and the effect of school characteristics on outcomes (Raudenbush, 2004). For example, Figure 9 indicates that not only do schools have differing degrees of shift between T1 and T2; there is also variation in how equitable the shifts are across the range of baseline scores within each school. In other words, in some schools baseline achievement is a much stronger determinant of T2 scores than in others. However, when dealing with three or more level hierarchical models, it is not always practical to allow the slopes to vary across the multiple levels since this can result in difficulties in achieving convergence of the parameter estimates.
Hierarchical linear models

Figure 9. Level 2 regression lines for individual schools with varying intercepts and slopes.

It is also possible to fix the intercepts in the model and allow the slopes to vary; for example, if all schools are known to have the same baseline achievement profile. However, this was not relevant to the current dataset, so is not shown here. The analyses in this chapter investigate the two possibilities illustrated above, in which intercepts are allowed to vary while slopes are fixed, or, where both slopes and intercepts are allowed to vary.

5.1.5. Steps in an HLM

The following steps are identical to those described in Chapter 4, illustrating the process required in conducting a regression analysis. Since HLM is simply a more complex representation of conventional regression, the steps hold across both paradigms. However, this added complexity does require additional explanation for some of these steps, described below. Again, this is not intended to be an exhaustive description of the HLM methodology since there are already several excellent text-books for this purpose (e.g., Gelman & Hill, 2007; Raudenbush, 2004), and should rather, provide sufficient background for the reader to be able to interpret the models presented later in this chapter.

• Statement of the research question
• Selection of potentially relevant variables
• Data Collection
5.1.5.1. Statement of the research question(s)

This is equally important in HLM as for conventional regression. As previously noted, an ill-posed research question can lead a researcher to select the wrong variables or predictors, incorrect model choice or analytic methodology, and ultimately, lead to a waste of considerable time and money. As noted in Chapter 4, financial constraints were of little concern in the current context since the data collection phase had already been completed, but it was still important that the research questions be appropriately operationalized to ensure that the analyses adequately addressed these questions.

5.1.5.2. Selection of potentially relevant variables

This step remains theory-driven, requiring the researcher to determine appropriate ways to measure relevant variables quantitatively, as described for conventional regression in Chapter 4. For example, if one is interested in the effect of SES on student outcomes, it would be necessary to determine how best to measure SES. What factors should be included (e.g., parental income, education level, books in the home)? Should the measure be categorical, or continuous? Opinions will differ as to which factors are most important, so the measure must be robust and grounded in the literature. In the HLM context, the level of analysis should also be identified for each variable. Using the simple two-level model in which students are nested in schools and T2 scores are predicted from baseline scores, SES could be an individual-level factor, a school-level factor, or both. If the researcher were only interested in the effect of attending a school with a particular socio-economic profile (i.e., the context effect), school-level aggregates would be
sufficient and individual-level (i.e., more invasive) information would be unnecessary, but this would preclude questions about the effect of individual-level SES on student outcomes.

5.1.5.3. Data collection

This step is the same for HLM as for conventional regression (see Section 4.1.3.3).

5.1.5.4. Model specification

In Chapter 4, it is noted that Samprit and colleagues (2006) argue that model specification involves the researcher using a combination of existing knowledge, as well as both objective and subjective judgments to make decisions about what form the model should take, indicating that the model specification is not fixed. Instead, it is a process whereby the researcher determines which variables should be retained in the model, meaning that the conceptualised model may need to be iteratively tested and updated during the analysis phase, resulting in a final model somewhat different from that initially theorised. This is also true for HLM models, but the process of model specification and testing differs somewhat.

As previously indicated, HLM attempts to estimate the relationship between one or more predictors on an outcome variable, while incorporating information about which level the information is derived from. Using our simple two-level example based on the random sample of Year 7 students, specification of an HLM for these data can be viewed as a two-stage series of analyses, with each stage corresponding to each level in the hierarchy. In the first stage (level 1), separate individual level regressions would be defined for each school, as expressed by the equation below:

$$Y_{ij} = \beta_{0j} + \beta_{1j} I_{ij} + \epsilon_{ij}$$  \hspace{1cm} (12)

The subscript $i$ represents individuals (students), while the subscript $j$ represents groups (schools). Therefore, $\beta_{0j}$ represents the school-specific intercept and $\beta_{1j}$ the school-specific slope of the predictor, $I$. The inclusion of both $i$ and $j$ subscripts indicates that the predictor, $I$, was
measured at the student level, hence $I_{ij}$ represents the student level predictor (T1 scores) for the $i^{th}$ student in the $j^{th}$ school, while $e_{ij}$ represents the residual error at the individual level. The residuals are assumed to be normally distributed with a mean of 0. Finally, $Y_{ij}$ represents the outcome variable (T2 score) for the $i^{th}$ student in the $j^{th}$ school. The equation follows the same form as for conventional regression, except for the addition of the subscripts, and indeed, the $i$ subscript can also be included in single level regression, but is usually excluded since the scores of individual students are not usually the level of interest. The subscript $j$ however, is not used in conventional regression, and is especially important in that it indicates that the parameter estimates can vary across different schools, as in Figure 8 where the intercepts varied, or as in Figure 9 in which both the intercepts and slopes were allowed to vary.

The second stage models the level 2-specific parameters shown in the equation above, as a function of level two (or higher) variables, as defined below:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}G_j + u_{0j}$$  \hspace{1cm} (13)

$$\beta_{1j} = \gamma_{10} + \gamma_{11}G_j + u_{1j}$$  \hspace{1cm} (14)

Remembering that $\beta_{0j}$ represents the school-specific intercept as shown in Figure 8, $G_j$ is the predictor variable at the school level, $\gamma_{00}$ is the aggregate intercept across all schools, and $\gamma_{01}$ is the effect of the individual school-level predictor on the school-specific intercepts. The school-specific slopes are captured by $\beta_{1j}$. $G_j$ is the predictor variable at the school level, $\gamma_{10}$ the common slope for all schools (derived from the aggregate of the student-level regression lines calculated for each school), with $\gamma_{11}$ the effect of the school-level predictor on the school-specific slopes. The gamma terms are fixed coefficients, applying to all students in the sample, regardless of school.

The residual terms represent the random effects part of the HLM, with $u_{0j}$ representing the random intercept component, and $u_{1j}$ the random slope component. In addition, these residual terms account for the reality that there will be a degree of sampling error at the group level, as
well as variables outside of the model that have an impact on group-specific outcomes. The residuals are assumed to be normally distributed with a mean of 0.

Both stages are shown below in a single equation, with the equations for the second stage substituted into the relevant terms for the stage one equation.

\[
Y_{ij} = \gamma_0 + \gamma_{01}I_{ij} + \gamma_{10}G_{ij} + \gamma_{11}G_{ij}I_{ij} + u_{0j} + u_{1j}I_{ij} + \epsilon_{ij}
\]

(15)

The equation includes the fixed effects at the school-level (\(\gamma_{01}\)) and student-level (\(\gamma_{10}\)), as well as their interaction (\(\gamma_{11}\)) on student outcomes (\(Y_{ij}\)). The random effects are modelled by the residual terms as described above. This general equation can be simplified by fixing certain aspects of the model. For example, the intercepts might be allowed to vary, but the slopes might be fixed, as in Figure 8. Alternatively, the equation might be extended to include more levels, with the possibility of fixed or random coefficients at and across each of the levels included in the model.

### 5.1.5.5. Selection of fitting method

In conventional regression, parameters are estimated using a single estimation method, such as the OLS procedure described in Chapter 4. In HLM however, the different model parameters; fixed and random effects, variance of the effects and residual variance, are estimated simultaneously, using different methods of estimation for each set of components.

**Fixed Effects**

The fixed effects in the equations in Section 5.1.5.4 are represented by \(\gamma\). These effects are level-2 effects (note that they were not in the level-1 equation). It is possible to estimate these fixed effects using the OLS estimation procedure, but since the precision of measurement is likely to vary across groups the use of OLS would violate the assumption of homoscedasticity and increase the risk of Type-I error. As a result, HLMs use Generalised Least Squares (GLS) estimates for these level-2 parameters, which results in a weighted level-2 regression, with weightings based on the degree of precision in the level-1 estimates. So, in the current context,
schools with less variation and more students would receive the greatest weight in the level-2 regression equation.

**Variance-Covariance Components**

The variance and covariance components in an HLM reflect the variance of the student level (level-1) residuals (i.e., the variance of $\varepsilon_{ij}$), and the variance and covariance of the school level residuals (level-2; i.e., the variance and covariance of $u_{0j}$ and $u_{1j}$). The data modelled using HLM techniques is typically unbalanced, in that the number of level-1 units (e.g., students) may vary across level-2 units (e.g., schools), and the pattern of level-1 effects may vary. As a result, iterative techniques such as the EM algorithm are used to provide maximum likelihood estimates of the variance-covariance components (Raudenbush, 2004).

**Level-1 Random Coefficients**

The level-1 random coefficients are represented in the equations described in Section 5.1.4.4 by $\beta_{0j}$ (school-specific intercepts) and $\beta_{1j}$ (school-specific slopes). This is frequently the level of focus for educational researchers (Hofmann, 1997) – for example, using our two-level example of 190 students in Year 7, school effectiveness could be investigated using the level-1 slopes for each school. This could also be estimated using the OLS procedure, with separate regressions for each school, and would provide reasonably precise estimates for schools with a large number of students represented in the analysis. However, these estimates become unstable as sample size declines (Burstein, 1980), meaning that smaller schools would not be evaluated appropriately. Recalling that HLM analysis can be viewed as a two-stage process, it can be seen that there are two estimates of the level-1 random coefficients; those derived using conventional OLS regression estimates within each unit/school (equation 12), and those derived from the level-2 equation – that is, the predicted values of $\beta_{0j}$ and $\beta_{1j}$ derived from data across similar units or schools (equations 13 and 14). As a result, for each unit (school), two estimates would be calculated for predicted intercept and slope values, one from each stage of the analysis.
Rather than attempting to determine which estimates of $\beta_0j$ and $\beta_1j$ are more appropriate, HLM utilises an empirical Bayes estimation method to combine the estimates into an optimally weighted estimate. These weighted estimates are based on the degree of precision of the OLS estimate and produce smaller mean square error terms than either procedure in isolation, ensuring that the Bayes estimate provides the best estimate of the level-1 random coefficients.

**Fitting/estimation options**

Most of the available software packages for HLM analysis allow certain aspects of the estimation procedures to be specified by the researcher. In MLwiN, the software used for the HLM analyses in this thesis, the researcher can choose three main estimation procedures; Iterative GLS (IGLS), Reweighted Iterative GLS (RIGLS), or Markov chain Monte Carlo (MCMC) simulation. IGLS and RIGLS are likelihood-based procedures, analogous to Maximum Likelihood and Restricted Maximum Likelihood, respectively. These procedures provide iterative point estimates for the parameters in the HLM until consecutive estimates are similar enough for the software to deem that convergence has been met (the convergence tolerance can be adjusted). RIGLS is preferable over IGLS when there are relatively few higher level units (e.g., if there are only fifteen schools). While IGLS and RIGLS were derived specifically for hierarchical modelling, MCMC methods have been applied to a wide range of applications and disciplines. MCMC estimation follows a Bayesian approach, which differs from likelihood or frequentist\(^\text{11}\) methods in that prior theory or knowledge is incorporated into the estimation. Each subsequent estimate then utilises the posterior theory/knowledge gleaned from the previous estimate as prior knowledge for the subsequent estimate, and so on (Browne & Rasbash, 2009).

A combination of RIGLS and MCMC estimation has been used in developing the models presented later in this chapter. As noted above, RIGLS is preferable when the number of higher level units is low (there are relatively few facilitators which were investigated as the highest level unit) and was used primarily for model building and to provide suitable priors for the MCMC

---
\(^{11}\) Frequentist methods base statistical inference on the frequencies apparent in data, while making no assumptions about the unknown values. This is the most common form of statistical inference in the social sciences.
estimation. MCMC estimation was used for all final (reported) estimates since it tends to out-
perform both likelihood methods (i.e., IGLS and RIGLS), especially when the number of groups is
small (even compared with RIGLS), there are multiple levels, and data are non-normal or missing
(Gill, 2002). All of these are considerations relevant to the LPDP dataset. Indeed, while all three
procedures perform relatively similarly in a two-level model, likelihood estimates suffered from
considerable bias compared with MCMC estimates for three-level models (Browne & Draper,
2006). Further, it is suggested that despite the considerable increase in computational time,
derivation of models using MCMC-based Bayesian techniques is a prudent approach since such
models typically outperform those based on likelihood estimates, and at worst, perform similarly.

5.1.5.6. Model fitting

Once the fitting method has been selected, the HLM can be conducted by fitting the model to
the available data using the selected method of estimation. The predictors investigated were the
same as those used in Chapter 4. As indicated above, RIGLS was used during preliminary model
testing (computation is much more rapid), followed by MCMC estimation to confirm results and
produce final parameter estimates. The parameter estimation and model fitting are automated
processes offered by a wide range of statistical software, including MLwiN\(^\text{12}\), HLM\(^\text{13}\), Mplus\(^\text{14}\), and
R\(^\text{15}\). Each of these software packages has different strengths, but MLwiN was used in this thesis
due to its extensive MCMC capabilities, combined with the ability to work with models with up to
five levels of hierarchy.

The standard procedure in building an HLM is very much an iterative process, starting with a
very simple model, then adding and testing the effect of predictors, usually one-by-one. As with
conventional regression, the order in which predictors are entered may influence whether the
predictor is deemed to have a significant effect on the outcome if the two predictors have a
degree of covariance. For example, in New Zealand, socio-economic status and ethnicity entered

\(^{12}\) http://www.bristol.ac.uk/cmm/software/mlwin/
\(^{13}\) http://www.ssicentral.com/hlm/
\(^{14}\) http://www.statmodel.com/
\(^{15}\) http://www.r-project.org/
independently tend to show a large effect on student outcomes, with more affluent students, and students of the dominant ethnicity (New Zealand European) typically achieving higher results than less affluent students, and those of other ethnicities. However, SES and ethnicity are related, and models incorporating both predictors typically show attenuation of the effect related to ethnicity when SES is included in the model. In conventional regression, the stepwise procedure allows researchers to automatically test and enter predictors, simply based on which predictors show the most reliable impacts on outcomes – which is not always the most desirable process. In HLM, stepwise procedures are not available, meaning the researcher must undertake a more systematic and theory-based process.

Typically researchers test the unconditional model first before adding predictors systematically to determine which are relevant to the hypotheses. The unconditional model includes no predictors and is essentially a variance partitioning model, assessing the proportion of the variance in outcomes explained at the different levels of the HLM, described earlier as the intra-class correlation (ICC). After this initial model, the researcher must make a number of decisions about which predictors to include, in which order, and at which of the higher levels each predictor should be allowed to vary (intercepts or slopes only, or both?; Raudenbush & Bryk, 2002). Generally this is not a set process, and the model is built iteratively, testing and investigating the effect of each predictor to determine how it should be included (or excluded) in the model. In the models developed in this chapter, the factors of interest were introduced one at a time. After building the complete model with all factors, covariates that did not reliably predict outcomes were then removed one at a time, and the effect on model fit considered.

There are several methods available to formalise the model-building process, including a multilevel extension of the $R^2$ calculation used in OLS regression. However, MLwiN calculates the deviance statistic, referred to as -2 log-likelihood, which provides an overall estimate of model fit. It is analogous to the sum of squared residuals in OLS regression, and indicates the overall degree to which the collected data differ from the estimation model developed. This provides a useful tool for comparing model performance with each additional parameter estimate (additional predictors, slopes or intercepts). Comparative model fit is assessed by calculating the difference
in deviance statistic between the two models; this follows the chi square distribution with degrees of freedom the same as the number of additional parameters being estimated. If the reduction in deviance is larger than the critical value for the chi square distribution, the model with the additional parameter(s) is deemed to be a superior fit, and inclusion of the parameter is warranted. This was the procedure used to determine whether covariates that had an unreliable (non-significant) effect on outcomes could be excluded with no significant effect on model fit.

5.1.5.7. Model validation and critique

As with conventional regression, validation of an HLM is a multi-step process in which the fitted model must be examined against statistical assumptions and theoretical interpretations, with the results of these examinations used to determine the performance and applicability of the model and guide any modification of earlier steps in the analysis (e.g., addition or removal of predictors, data transformation, estimation techniques and so on). The assumptions made while determining the form of the model (i.e., which predictors, estimation method, linearity and so on) must hold if the model is to be valid, so adequate planning and assessment of the form the model should take, helps to expedite the validation process.

In the HLM context, after determining which predictors should be included in the model (usually determined by whether the predictor has a sufficiently reliable effect on the outcome variable), and whether or not to allow the slopes / intercepts for each predictor to vary across higher levels in the model (assessed by assessing the degree of variance across groups), validation is similar to the process used in conventional regression. Indeed, the first step is usually to assess the residuals to check the model specification, by ensuring that the residuals are normally distributed and homoscedastic. In HLM however, there are two (or more, depending on the number of levels) residuals. In order to assess the model, the residuals should ideally be considered separately, but this is not possible for higher-level residuals (only those at level one; Snijders & Bosker, 2011). Therefore, it is sufficient to assess the residuals of each group of students separately using the level-one OLS estimates in the same way and against the same
assumptions as in conventional regression, but the residuals at higher levels of analysis are somewhat more complex.

At higher levels of analysis, raw residuals must be multiplied by a shrinkage factor, which shrinks the residuals as a function of variance at each level and group-level sample size, as shown below.

\[
\hat{u}_{bj} = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_e / n_j} r_{*,j}
\]

The \(\hat{u}_{bj}\) term is the residual seen in the equations described in Section 5.1.5.4, which allowed the school-level mean to be adjusted from the overall mean. The addition of the ‘hat’ simply indicates that this is the estimated residual. The raw residual is represented as \(r_{*,j}\), the mean individual level residual for the \(j\)th school. The \(\sigma^2\) terms represent the variance at each level, with the \(u\) subscript referring to level-two variance, and the \(e\) subscript representing level-one variance. Finally, the sample size of the school is represented by \(n_j\).

In essence, the shrinkage equation allows the level-two estimates to be adjusted for the degree of precision. As a result, the shrinkage adjustment is largest when there are relatively few students in each group, when there is a lot of level one variance, and when the level two variance is small. Once the level-two adjusted residuals have been calculated, they can be used to check model assumptions, such as the normality assumption (which applies to each level). As with conventional regression, normality can be assessed using a residual plot, or via the Kolmogorov-Smirnov test to determine whether the standardised residuals follow a normal distribution with a mean of zero. The standardised residuals should also have minimal correlation with each of the predictors, with a scatter plot of the residuals against the predictors showing no detectable patterns. Graphical methods are particularly useful for confirming the performance and validity of regression models since graphs can provide insight into features of the data that are extremely difficult to identify using formal diagnostics (Huber, 1991).
5.1.5.8. Usage and reporting of the derived model

Ultimately the intention of HLM analysis is the same as for conventional regression. That is, to present a model that describes the relationship between a series of predictors and a particular outcome of interest. This can then be used to inform policy, measure the effects of changes in policy, assess relationships and inter-relationships between variables, or predict likely outcomes for individual students and use these predictions to provide targeted teaching. In order to ensure that the conclusions drawn from the developed models remain valid, it is essential that the regression assumptions be assessed, and that conclusions and predictions remain within the range of data, and level of analysis used in developing the original model.

5.1.6. Additional Considerations

The steps described in the previous section provide an introduction to the basics of HLM and how these relate to conventional regression as well as to the analyses presented in this chapter. However, there remain some additional considerations relevant to this thesis that were not described in these steps, and therefore, are described below.

5.1.6.1. Longitudinal analysis

The analysis of repeated measures data using HLM techniques is advantageous since attrition is less problematic compared with other modelling procedures (e.g., Field, 2009; Hox, 2000; Twisk, 2006). Participants with missing data at one or more time point will still be included in the analyses but parameter estimates for these individuals are weighted to account for the lower reliability in these estimates. Using conventional techniques, individuals with missing outcome data would be excluded.

This is achieved by recognising the nested nature of longitudinal data, whereby each observation is a level-one unit nested within the individual at level-two (Bryk & Raudenbush, 1987). For example, the two-level model discussed in the previous section, modelling T2 achievement as the outcome measure, with students nested within schools would be re-specified...
as a three-level model where subsequent test scores are level-one observations nested within students at level-two and schools at level-three.

5.1.6.2. Centring

In conventional regression, centring is largely a question of convenience since centring does not change the parameter estimates, and merely changes the intercept parameter, often making interpretation easier. In HLM however, since the intercepts and slopes at level-one are used as outcomes in the level-two equation (thereby meaning that transformations that affect the intercept will have an impact on the level-two parameter estimates), centring is a real issue, and can affect interpretation of the slopes (Kreft, de Leeuw, & Aiken, 1995; Longford, 1989; Plewis, 1989; Raudenbush, 1989). In a practical sense, in the multilevel context centring may also reduce the likelihood of estimation errors when using IGLS or RIGLS as the estimation method (Rasbash et al., 2009).

There are three main centring options; grand mean centring, group mean centring or simply using the variable in its raw (non-centred) form. With grand mean centring, the overall mean for the whole cohort is subtracted from each student’s score on the predictor, while group mean centring refers to the subtraction of the group-specific mean from the individual’s score. Models produced using grand mean centring are equivalent to those produced using the raw data, but are argued to be more desirable since grand mean centred models are more interpretable and have fewer issues with multi-collinearity (Kreft et al., 1995). Conversely, models derived using group mean centred data are not equivalent to those produced using grand mean or raw form centring. Despite this, all three options are statistically defensible, and the decision about whether and how to centre data must relate to the research question, with interpretation taking the centring method into account (Hofmann & Gavin, 1998).
5.2. Analysis

The following analyses explore the relationship between the factors identified as of primary interest throughout the thesis (i.e., gender, decile, and ethnicity), as well as the additional school-level factors explored in the single-level regression models (i.e., school size, type and proportion of minority students), using HLM. This allows direct comparison of the results obtained using the different methodologies to analyse the same data, and also provides a clearer picture of student progress over the intervention period. MLwiN version 2.26 has been used, with MCMC as the main estimation method, since it tends to out-perform likelihood methods (such as maximum likelihood, ML) when the number of groups is small (in this case, there are relatively few facilitators), and when data are non-normal or missing (Gill, 2002), both of which are considerations here. Indeed, while likelihood techniques and MCMC estimation perform relatively similarly in a two-level model, considerable bias is evident in the likelihood estimates for three-level models compared with those derived using MCMC (Browne & Draper, 2006). Further, it is suggested that despite the increase in computational time, derivation of models using MCMC-based Bayesian techniques is a prudent approach since such models typically outperform those based on likelihood estimates, and at worst, perform similarly. RIGLS was used to provide the start-point for the MCMC estimation, and these results are also presented to provide an indication of the magnitude of differences that may arise due to estimation method. A burn-in of 500 with simulation of 10,000 iterations was used in all cases.

The LPDP data were conceptualised as five-level longitudinal models, with observations on the dependent variable (asTTle score versus normative expectation) as the level-1 unit. As noted in section 5.1.5.1, this is advantageous since it allows students with missing data to be included in the analysis. As with any large-scale intervention in an educational setting, not all students could be tested at each time point. Whereas single level techniques (i.e., MANOVA, multiple regression and effect sizes) typically use listwise deletion of all cases where a single observation is missing, multilevel models do not assume that there are an equal number of observations for each student (Hox, 2000), instead incorporating the uncertainty inherent with missing values into the model. There were very few cases (<0.1%) where explanatory variables were unavailable.
since these data were provided by schools’ student-management systems. Observations were nested within students (level-2), students within year-groups (level-3) clustered within schools (level-4). Since each facilitator worked with a small group of schools in a specific region, the clustering of schools within facilitation/regional groups was also investigated as a level-5 unit. Teachers and/or class are another potential source of (cross-classified) clustering, but due to anonymization of the data this could not be analysed. However, since many primary schools in New Zealand are relatively small, classes in the same year level frequently work closely in teaching teams, making this a reasonable source of clustering in the data. On average there were ~40 students per year level in each school for both reading and writing. Due to the nature of HLM the number of cases at each level becomes a more prominent feature of the analysis than for other methods, where the number of students is typically the focus. The number of cases at each level is shown in Table 25. Note that the number of students differs from previous analyses; as stated in this chapter’s introductory section one of the advantages of HLM is that students do not need to have observations for all time points. As a result, all students for whom at least one assessment was obtained are included in the models.

As noted at the beginning of this section, this chapter investigates the project effect of the factors identified to be of primary interest throughout the thesis (i.e., gender, decile, and ethnicity), along with the additional school-level factors explored in the single-level regression models (i.e., school size, type and proportion of minority students) using HLM. However, because HLM procedures partition variance according to level it is necessary to identify the level at which each
factor is entered into the model, so a synopsis of the variables and levels is included below. The implications of centring are generally an important consideration when using HLM, so the centring procedure used is also identified for each variable. As discussed in section 5.1.6.2, the major decision is whether to centre based on the grand mean (CGM), or centre within cluster (CWC), since CGM gives the same parameter estimates as using the raw scores (Enders & Tofighi, 2007). CWC is usually not used for higher level units since there is either no further clustering, or the relationship is not of interest. For example, while school size is of interest with respect to the student-level outcomes, cross-level interactions between school size and facilitator (the higher level) is not of interest. At the lower levels, the variables are all categorical, so variables have been transformed to ensure that zero is the reference group, and one is the comparison group (for binary variables). While it is possible to centre dummy variables in HLM (and sometimes useful to do so), the use of binary dummies with zero as the reference category is useful for interpretability.

Level-1: Observations. Since progress over time is of interest, time is added as a level-1 covariate so that the model can represent growth as a straight line with a non-zero slope. The time covariate is entered as a scale variable, with baseline scores set to zero, EOY1 to one, and EOY2 to two. Time is also entered as a first-order cross-level interaction term, so that differential rates of progress among subgroups can be investigated.

Level-2: Students. The variables at this level were entered in the same way as for conventional regression. Explicitly, student ethnicity was dummy coded. There were four ethnic categories; NZ European, Māori, Pasifika and ‘other’. Students in the ‘other’ category were predominantly of Asian descent, but also include a smaller proportion of Middle Eastern and Continental European students. Students in this category typically have an achievement distribution similar to that of NZ European students; therefore, as with the conventional regression analyses, these students were included with NZ European students as the reference category. Since this left three categories, two covariates were added to the model (n-1). Year level was also dummy coded, with students in Year 4 treated as the reference category, while gender was included as a binary dummy with male students as the reference.
Level-3: Year/Grade. As indicated earlier, it was theorised that there would be some clustering evident at the year-level due to the nature of NZ schools, hence the inclusion of this in the hierarchical structure of the model. However, there were no covariates at this level.

Level-4: School. School size was entered as the actual roll size for each school to ensure maximum variance was retained, and the percentage of minority students attending each school was also investigated. Minority was defined as students identifying as Māori and/or Pasifika, and does not include students in the 'other' category since this latter category does not suffer from an achievement gap compared with students identifying as NZ European. Both of these variables were centred around the school-level grand mean (i.e., the average school size or proportion of minority students). School decile was included as a categorical variable at this level, with low decile treated as the reference category and mid and high decile schools the comparison groups.

Level-5: Facilitator/regional level. Facilitator demographics and experience were possible covariates, but given the relatively small number of facilitators, it was considered unlikely that any differences would be sufficiently reliable.

Interactions and moderating variables were also examined to investigate moderating effects. For example, is the effect of being in a low decile school different for males than it is for females? Does being in a low decile school affect students of a particular ethnicity more than those of a different ethnicity? All interactions that made theoretical and statistical sense were examined, but no reliable interaction effects were identified, so these are not discussed further.

The random intercept model is shown below:

\[ y_{ijklm} = \beta_0 + e_{ijklm} \]  

(17)

where \( y \) is the asTTle achievement score against normative expectation; the subscripts identify levels one (i) through five (m); \( \beta_0 \) the intercept; and \( e \) the residual error. The unconditional models were identical for both reading and writing. The models were then built iteratively, investigating which predictors explained a significant amount of the variance in achievement scores, with each covariate being examined separately prior to inclusion in more complex models. Note that the subscript (i) is missing from the intercept in this equation; this is because
longitudinal multilevel models cannot have random intercepts at level-1 since such models represent the student-level growth as a flat line with a slope of zero (Peugh, 2010). In other words, the prediction about each individual’s performance is derived by the aggregate information from each of the individual test scores (observations) for each student, with conclusions about time reliant on the inclusion of a time covariate. As noted above, time has been centred by subtracting one from each time point, so that baseline (time one) is coded as 0. This allows the lower order coefficients to be interpreted as baseline differences where these are also centred so that 0 is the reference category. The first order time covariates reflect progress attributable to the project, since the normative expectations have been subtracted.

For each cohort, unconditional models were estimated initially. These models allow the proportion of variance in student outcomes explained at each level to be calculated, without the inclusion of any explanatory factors. After fitting the unconditional models, the procedures outlined in section 5.1.4.6 were followed, with covariates added separately to examine their individual contribution before developing more complex models. The interaction of time with each covariate, allowing different progress trajectories for each subgroup, was captured by adding the cross-level first order interaction previously mentioned (effectively a multiplicative term that allows the contribution of each covariate to progress to be assessed). All factors were then entered into an overall model (both the lower order and time interaction terms), with the least significant factors removed one by one then assessed for loss of variance explained. This is assessed by comparing the change in deviance against a chi-square distribution, as explained in subsection 5.1.5.6. The final models include only predictors that continued to explain a significant source of variance after the inclusion of other variables.

Note also that it is possible to fix parameter estimates across all levels of the model, or to allow the estimates to vary at selected levels. Allowing parameter estimates to vary across levels can improve model fit but also increases complexity, so initially factors were added as fixed factors, and those that were both significant and explained sufficient variance were then investigated further (when this made sense to do so theoretically) to determine the value of allowing the estimates to vary across levels. Interpretation of the lower order covariates is
relatively simple, and can be made in much the same way as for conventional regression, whereby the specific parameter estimate can be interpreted as the cost or benefit of being in the particular subgroup. However, specification of the HLM as a longitudinal growth model has allowed the relative differences in achievement and progress to be modelled within a single model. Since time has been centred so that baseline is coded as 0, the lower order coefficients can be interpreted as baseline differences. Interpretation of relative progress is also relatively simple, though the addition of the higher order time interaction terms adds considerable complexity to the overall model interpretation.

As with the analyses in the previous chapters, the standardisation procedure whereby the expected asTTle score for each student’s year level was subtracted from his/her score allows the coefficients relating to progress over time to be reflective of progress after removing maturational effects, but note that the coefficients reflect the predicted change for each year of the project. That is, the coefficients should be doubled to determine predicted progress over the two years of the project (i.e., if comparing against the regression coefficients in chapter 4). The time interaction terms allow pair-wise comparison of the shift for that particular subgroup with its initial achievement. For example, if females have initially higher achievement, it is possible to determine the extent to which this changes over the course of the project by looking at the unstandardized covariate of gender by time. However, while comparative interpretations such as the above are relatively simple, interpretations about the project value are more complex, and must take into account the aggregate information provided by the whole model rather than simply looking at individual coefficients. After presenting and explaining each of the models developed, the facilitator effect is considered by investigating the residuals at the facilitator level, for comparison against the results of the value-added investigations outlined in Chapter 4.

5.2.1. Reading Focus – 2006-2007 Cohort

The proportion of variance explained at each level is shown below in Table 26. The distribution of variance was similar for both estimation methods at most levels, but the MCMC
procedure yielded a lower variance estimate at the facilitator level, and showed lower model deviance. The variance explained at the facilitator level was non-significant irrespective of the estimation procedure, but since there are no negative effects associated with the inclusion of levels where variance explained is low, the level was retained for consistency, and to allow comparisons of the value-added facilitator effect in subsequent analysis. The variance of the residual term, which includes differences over time (as well as other, unknown sources of variance), was both large and significant. Student-level variance was also large, with almost half of the variance in asTTle scores explained at the student level. Year level was a small, but still significant source of variance. School level variance was a little lower than that described elsewhere (e.g., Lietz, 2009 indicates that typically 17% of the variance is at the school level in NZ, a relatively low figure internationally), but this is largely because the facilitator level is usually not included. If the variance at the facilitator level is redistributed to a four-level model, the variance explained at the fourth-level (which would be defined as the school-level) is very close to 17%.

Table 26. Unconditional models comparing variance proportion at each level by estimation procedure.

<table>
<thead>
<tr>
<th></th>
<th>RIGLS</th>
<th>MCMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facilitator</td>
<td>5.5% (n.s.)</td>
<td>3.7% (n.s.)</td>
</tr>
<tr>
<td>School</td>
<td>12.1%***</td>
<td>15.1%***</td>
</tr>
<tr>
<td>Year cluster</td>
<td>2.8%***</td>
<td>2.7%***</td>
</tr>
<tr>
<td>Student</td>
<td>46.2%***</td>
<td>45.6%***</td>
</tr>
<tr>
<td>Time (+ residual)</td>
<td>33.3%***</td>
<td>32.9%***</td>
</tr>
<tr>
<td>Deviance</td>
<td>175840.7</td>
<td>166467.1</td>
</tr>
</tbody>
</table>

* p < 0.05; ** p < 0.01; *** p < 0.001

To provide a baseline picture of the overall project effectiveness, the unconditional model with time as a covariate, is shown in the equations below. These equations indicate that baseline achievement among students in this cohort was approximately 16 points higher than the asTTle normative expectations, and that progress was around nine points higher per year (overall
progress SDU = 0.2). The unexplained terms in each equation represent the error, or residual terms, for each level.

\[
ScoreVsExpectation = \beta_{0ijklm} + 9.32(0.52)(Time - 1)_{ijklm} + e_{ijklm}
\]  

(18)

Where: \( \beta_{0ijklm} = 16.12(5.57) + g_{0lm} + f_{0lm} + v_{0klm} + u_{ijklm} \)

The parameter estimates and their associated standard errors for the final model are shown for each factor in Table 27. Two lower order covariates that did not explain a significant amount of the variation in student scores include Year 5 and mid decile. That is, students in these groups had baseline scores no different from the relevant reference group. Exclusion of these terms from the model simply means that the coefficients estimated reflect differences against an aggregated reference group, whereby the Year 6 and 7 terms are referenced against students in Years 4 and 5, whose scores were not significantly different from each other at baseline. In addition, there were initially no differences in achievement by school type, with similar results among students attending full primary, contributing and intermediate schools. There were also four first order time covariates that were non-significant, with progress among Pasifika, female, Year 5 students and those attending full primary schools, no different than for the relevant reference group. That is, the gender gap remained similar over the course of the project, progress rates were similar for students in Years 4 and 5, Pasifika students continued to have achievement almost half a standard deviation below students of NZ European and other ethnicities, and students in full primary and contributing schools progressing at similar rates.
Table 27. Final HLM showing parameter estimates by estimation method for 2006-2007 reading cohort.

<table>
<thead>
<tr>
<th></th>
<th>RIGLS estimates</th>
<th>MCMC estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unstandardized</td>
<td>Standardized</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td>Intercept</td>
<td>18.33</td>
<td>4.79</td>
</tr>
<tr>
<td>Year 6</td>
<td>-15.42</td>
<td>3.63</td>
</tr>
<tr>
<td>Year 7</td>
<td>-11.99</td>
<td>5.76</td>
</tr>
<tr>
<td>Female</td>
<td>18.78</td>
<td>1.69</td>
</tr>
<tr>
<td>Pasifika</td>
<td>-46.07</td>
<td>4.34</td>
</tr>
<tr>
<td>Māori</td>
<td>-30.85</td>
<td>2.61</td>
</tr>
<tr>
<td>School Size</td>
<td>-0.053</td>
<td>0.019</td>
</tr>
<tr>
<td>Minority Proportion</td>
<td>-0.361</td>
<td>.159</td>
</tr>
<tr>
<td>High Decile</td>
<td>35.21</td>
<td>8.51</td>
</tr>
<tr>
<td>Year 6 * time</td>
<td>11.79</td>
<td>1.53</td>
</tr>
<tr>
<td>Year 7 * time</td>
<td>25.85</td>
<td>2.45</td>
</tr>
<tr>
<td>Māori * time</td>
<td>-4.25</td>
<td>2.33</td>
</tr>
<tr>
<td>School Size * time</td>
<td>0.03</td>
<td>0.003</td>
</tr>
<tr>
<td>Minority * time</td>
<td>0.315</td>
<td>0.038</td>
</tr>
<tr>
<td>Intermediate * time</td>
<td>15.34</td>
<td>2.77</td>
</tr>
<tr>
<td>Mid Decile * time</td>
<td>8.28</td>
<td>1.64</td>
</tr>
<tr>
<td>High Decile * time</td>
<td>17.38</td>
<td>1.97</td>
</tr>
</tbody>
</table>

*The interaction effect of Māori over time has been allowed to vary by school

** This factor is non-significant when allowed to vary across levels, but is of value to the model since model deviance is reduced considerably ($X^2 > 60, \ p < .001$).

Students in Years 6 and 7 had marginally lower initial achievement than the younger students (SDU = .1 to .15); the difference for students in Year 7 was non-significant when estimated using MCMC. This gap reversed over the course of the project, with higher progress rates among the
older students suggesting a 40 point advantage among Year 7 students by EOY2 (progress was
~25 points higher per year, but ~10 points lower initially). The marginally higher rate of progress
among Year 6 students resulted in EOY2 scores similar to those in Years 4 and 5. While Pasifika
students progressed at approximately the same rate as NZ European and students of ‘other’
etnicities, NZ Māori made marginally less progress. This was significant when fixed across
levels, but the estimate of decline was halved, and non-significant, when allowed to vary at the
school-level. Nevertheless, the factor was retained since it explained considerable additional
variance. This indicates that there was significant variation in the degree to which individual
schools supported progress among Māori.

Those in low and mid decile schools also had similar baseline scores (hence the exclusion of
the covariate for mid decile), while students in high decile schools typically had baseline
achievement around 35 points higher. Progress among students in mid-decile schools was
marginally higher than for those in low decile schools however (SDU = .15). The gap for students
attending high decile schools widened over the course of the project, with much faster progress
(SDU = .7) resulting in an achievement difference of a full standard deviation between students
attending low versus those attending high decile schools. As previously noted, school type was of
little significance at baseline, with students of all school types sharing a similar achievement
pattern. However, progress among students attending intermediate schools was higher, leading
to an EOY2 advantage of approximately one-third of a standard deviation.

School size, and the proportion of minority students attending a school, had a small initial
effect on achievement, with students in larger schools, and those in schools with a higher
proportion of NZ Māori and Pasifika students typically having lower baseline scores. These
differences were small however (the difference by proportion of minority students is non-
significant when estimated with MCMC methods), with a standardised mean difference of around
one-fifth to one-quarter of a standard deviation between students attending schools at either
extreme. The relative advantage of being in a smaller school diminished over the course of the
project, with students in schools of all sizes achieving similar EOY2 results. Students attending
schools with a higher proportion of Pasifika and Māori students typically had higher rates of
progress, this increased rate of progress (approximate SDU = .5) was large enough that the gap was reversed by EOY2, such that the initial disadvantage of attending schools with a higher proportion of NZ Māori and Pasifika students became an advantage of similar magnitude. This needs to be taken into account when assessing the progress rates of students of specific ethnicities. While the progress rates for Māori and Pasifika appear insufficient to close the achievement gap, many Māori, and especially Pasifika, attend schools where there are large proportions of minority students. Therefore, for many students of Māori and Pacific descent, these gaps do diminish, but the school-level effects are of greater importance.

At first glance, it appears that overall progress has been negative, with a loss of approximately one-third of a standard deviation relative to the normative sample over the course of the project. However, due to the sometimes considerable differences in rates of progress among specific subgroups, the overall progress rates are somewhat difficult to glean from the overall model. Instead, this progress rate reflects progress when all of the covariates are at zero; that is, for NZ European/Other students of either gender in Year 4 or 5, attending a low/mid decile full primary or contributing school, of average size with an average proportion of Māori and Pasifika students. It is possible to manipulate each individual covariate to determine which groups had rates of progress higher than the normative expectation. For example, it is clear that the same set of circumstances would no longer result in a decline in achievement over time if the student were in Year 7, where progress was 26 points higher than the reference group, and therefore ten points higher for each year, than the normative sample (obtained by adding the Year 7 ‘advantage’ to the overall ‘decline’). However, to get an overall picture of the average project-level progress, the initial model discussed above, where time is the only covariate, should be used since this provides the ‘average’ achievement and progress among all students in the cohort.

Table 28 indicates the proportion of remaining variance and variance explained by estimation method. The remaining unexplained variance was almost entirely at the student and time levels – indeed, a relatively small proportion of variance was explained at these levels. Less than 5% of the variance at the student level was explained, while the proportion of variance in progress
explained was modest. There was a considerable difference in the estimated proportions of year-
level variance explained by estimation method, with MCMC indicating only 5% explained, while
RIGLS estimated three times as much variance explained. More than half of the school-level
variance was explained, and almost all of the variance at the facilitator level. Since there were no
facilitator factors included, this suggests that the variance at the facilitator level was explained by
school-level factors, suggesting little (if any) variance in outcomes by facilitator. This is examined
further in the value added section.

Table 28. Remaining variance and variance explained by estimation method.

<table>
<thead>
<tr>
<th></th>
<th>Proportion of remaining variance</th>
<th>Proportion of variance explained</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RIGLS</td>
<td>MCMC</td>
</tr>
<tr>
<td>Facilitator</td>
<td>0.8%</td>
<td>0.5%</td>
</tr>
<tr>
<td>School</td>
<td>7.1%***</td>
<td>7.5%***</td>
</tr>
<tr>
<td>Year cluster</td>
<td>3.1%***</td>
<td>3.3%***</td>
</tr>
<tr>
<td>Student</td>
<td>57.8%***</td>
<td>55.7%***</td>
</tr>
<tr>
<td>Time (+ residual)</td>
<td>34.2%***</td>
<td>33%***</td>
</tr>
<tr>
<td>Reduction in Deviance</td>
<td>3528.1***</td>
<td>4421.2***</td>
</tr>
</tbody>
</table>

* p < 0.05; ** p < 0.01; *** p < 0.001

Overall, the results were comparable across estimation methods, with very similar point
estimates for all covariates. However, there were two instances in which the RIGLS estimation
indicated a significant effect on student outcomes for a covariate, which was not supported by the
MCMC estimation. In addition, the allocation of variance at each level differed somewhat by
procedure. In some cases these differences were substantial, with the RIGLS suggesting
considerable variance explained at the Year level, while the MCMC estimation indicated the
explanation of variance at this level was of marginal significance. The results were also broadly
comparable to the estimates derived using single level regression in section 4.2.2.1. Both models
were relatively complex, and generally the same factors were identified as having a significant
effect on student outcomes. However, there were some interesting differences of note; for
example, the gain among Māori was estimated to be fairly close to the overall project using HLM,
while the regression model suggested gains were much lower (SDU = -.25). This difference is largely a result of allowing the variance in gains among Māori across schools to be modelled, more closely reflecting the data. Other differences are indicated and discussed in Chapter 6.

5.2.2. Reading Focus – 2008-2009 Cohort

The proportion of variance explained at each level is shown below in Table 29. Again, the distribution of variance was similar for both estimation methods at most levels, but as with the previous reading cohort, the MCMC procedure yielded a lower variance estimate at the facilitator level, and showed lower model deviance. The variance explained at the facilitator level was non-significant for both estimation procedures, but the level was retained to allow comparisons of the value-added facilitator effect in subsequent analysis. The variance at the school-level was non-significant when using RIGLS estimation, but significant when using MCMC. The combined variance of school and facilitator was a little lower than the previous reading cohort (13-15%), and slightly lower than Lietz’s figure of 17% for NZ schools. However, this cohort is comparatively small, so it may be that the lower variability at the higher levels is simply due to the differences in intake – schools that opted to use asTTle shared some similarities compared with the schools in this cohort that opted to use the Progressive Achievement Test (Reid & Elley, 1991). Year level was a small, but still significant source of variance. Much of the variance in asTTle scores was at the student level, with almost as much variation over time, though the ‘time’ level incorporates the majority of the error so this is unsurprising.
Table 29. **Unconditional models comparing variance proportion at each level by estimation procedure.**

<table>
<thead>
<tr>
<th>Level</th>
<th>RIGLS</th>
<th>MCMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facilitator</td>
<td>4.4% (n.s.)</td>
<td>2.4% (n.s.)</td>
</tr>
<tr>
<td>School</td>
<td>9.2% (n.s.)</td>
<td>13.1%*</td>
</tr>
<tr>
<td>Year cluster</td>
<td>5.7%***</td>
<td>5.8%***</td>
</tr>
<tr>
<td>Student</td>
<td>44.3%***</td>
<td>41.2%***</td>
</tr>
<tr>
<td>Time (+ residual)</td>
<td>38.3%***</td>
<td>37.4%***</td>
</tr>
<tr>
<td>Deviance</td>
<td>79759.7</td>
<td>75828.5</td>
</tr>
</tbody>
</table>

*p < 0.05; ** p < 0.01; *** p < 0.001

To provide an initial estimate of the overall project effectiveness, the unconditional model is shown in the equations below, with time as a covariate. These equations indicate that baseline achievement among students in this cohort was approximately 13 points lower than the asTTle normative expectations, and that progress was around sixteen points higher per year (overall progress SDU = 0.33). The unexplained terms in each equation represent the error, or residual terms, for each level.

\[
ScoreVsExpectation = \beta_{0,jklm} + 16.5(0.87)(Time - 1)_{ijklm} + e_{ijklm}
\]  

(19)

Where: \(\beta_{0,jklm} = -13.21(9.58) + g_{0m} + f_{0lm} + v_{0klm} + u_{0,jklm}\)

The parameter estimates and their associated standard errors for the final model are shown for each factor in Table 30. Specifically, there were no baseline differences by school type, size or proportion of minority students attending. The rate of progress among students attending schools with different proportions of minority students was also similar, but was greater in smaller schools. In addition, students attending full primary schools also progressed more rapidly (SDU = .25). There were no differences between low and mid decile schools at baseline, but students attending high decile schools had initial achievement approximately half a standard deviation higher than students in other deciles. This gap also widened over time, with an additional gain among students in high decile schools of one third of a standard deviation relative to students in
low decile schools. Students in mid-decile schools also progressed more rapidly than those in low decile schools (SDU = .2).

Table 30. Final HLM showing parameter estimates by estimation method for 2008-2009 reading cohort.

<table>
<thead>
<tr>
<th></th>
<th>RIGLS estimates</th>
<th></th>
<th>MCMC estimates</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unstandardized Coefficients</td>
<td>Standardized Coefficients</td>
<td>Unstandardized Coefficients</td>
<td>Standardized Coefficients</td>
</tr>
<tr>
<td>Intercept</td>
<td>-25.12</td>
<td>7.53</td>
<td>-25.63</td>
<td>8.43</td>
</tr>
<tr>
<td>Female</td>
<td>21.86</td>
<td>2.54</td>
<td>21.96</td>
<td>2.57</td>
</tr>
<tr>
<td>Māori</td>
<td>-19.58</td>
<td>3.62</td>
<td>-19.4</td>
<td>3.56</td>
</tr>
<tr>
<td>Pasifika</td>
<td>-28.89</td>
<td>4.43</td>
<td>-29.08</td>
<td>2.57</td>
</tr>
<tr>
<td>High Decile</td>
<td>48.49</td>
<td>16.32</td>
<td>47.39</td>
<td>17.08</td>
</tr>
<tr>
<td>Time</td>
<td>-8.79</td>
<td>3.07</td>
<td>-8.86</td>
<td>3.11</td>
</tr>
<tr>
<td>Year 5 * time</td>
<td>6.58</td>
<td>2.02</td>
<td>6.64</td>
<td>2.03</td>
</tr>
<tr>
<td>Year 6 * time</td>
<td>21.63</td>
<td>2.36</td>
<td>21.56</td>
<td>2.39</td>
</tr>
<tr>
<td>Year 7 * time</td>
<td>37.93</td>
<td>2.53</td>
<td>37.93</td>
<td>2.53</td>
</tr>
<tr>
<td>Mid Decile * time</td>
<td>9.98</td>
<td>3.01</td>
<td>10.05</td>
<td>2.99</td>
</tr>
<tr>
<td>High Decile * time</td>
<td>16.63</td>
<td>3.01</td>
<td>16.66</td>
<td>3.02</td>
</tr>
<tr>
<td>School size * time</td>
<td>-.035</td>
<td>.005</td>
<td>-.035</td>
<td>.005</td>
</tr>
<tr>
<td>Full Primary * time</td>
<td>13.39</td>
<td>2.91</td>
<td>13.45</td>
<td>2.95</td>
</tr>
</tbody>
</table>

Female students had initial achievement 22 points higher than for males, and this remained static over the course of the project. There were no differences in baseline achievement by year level, but older students gained more over the course of the project. The increased rate of progress among Year 5 students was of marginal practical significance (SDU = .1), but moderate among Year 6 students (SDU = .4) and large among students in Year 7 (SDU = .75). Māori and Pasifika students had initial achievement lower than students of other ethnicities (including NZ...
European), and this gap remained similar throughout the duration of the project. All factors were fixed in this model, since no significant improvements in model fit that were theoretically meaningful were identified. Again, the coefficient for progress over time is negative. However, as with the previous cohort, there were considerable differences in the rates of progress among specific subgroups, making the overall progress rates somewhat difficult to determine. Within this model, the coefficient for time reflects the progress among NZ European/Other students of either gender in Year 4, attending a low decile contributing school, of average size. It can be seen that students in Years 6 and 7 typically had rates of progress higher than expectation. Note that since the progress effects are cumulative, the easiest way to determine the average project-level progress is via the initial model discussed above, where time is the only covariate, in which the estimated added value of the project was a third of a standard deviation.

Table 31 indicates the proportion of remaining variance, as well as the variance explained, according to the two estimation methods employed. As with the previous reading cohort, the remaining unexplained variance was almost entirely at the student and time levels and a relatively small proportion of variance was explained at these levels. Less than 1% of the variance at the student level was explained, while the proportion of variance in progress explained was modest. There was a considerable difference in the estimated proportions of variance explained by estimation method for the higher levels, with RIGLS indicating that very little of the school-level variance was explained (5%), and all of the variance at the facilitator level was explained. Conversely, the MCMC estimation produced more plausible results, indicating that much of the variance (but not all) at the facilitator level was explained and a modest proportion at the school level. These differences, while considerable, are not unexpected since there were fewer schools and facilitators and stability of estimates is somewhat less at higher levels. The minimal variance remaining at the facilitator level calls into question whether there is any meaningful variance in outcomes by facilitator. This is examined further in the value added section.
The results were remarkably stable across estimation methods, with very similar point estimates for all covariates, and no differences in substantive interpretations. However, the allocation of variance at each level differed somewhat by procedure and in some cases these differences were quite large. The RIGLS estimates indicated minimal variance explained at the school level, while the MCMC estimation indicated more than a quarter of the variance at this level was explained. The results were also broadly comparable to the estimates derived using single level regression in section 4.2.2.2. Both models were much simpler than the models for the previous reading cohort, and generally the same factors were identified as having a significant effect on student outcomes. The estimates of the additional gain among students in mid and high decile schools were equivalent for both single level regression and HLM. Again however, there were some interesting differences of note; for example, the gain among students of different ethnicities was indicated to be equivalent to the overall cohort in both cases, but while the single level models suggested a negative effect of being in a school with a high proportion of minority students, this was not supported by the HLM analyses. Additional differences are discussed in Chapter 6.

5.2.3. Writing Focus – 2006-2007 Cohort

The proportion of variance explained for each level in the model is shown below in Table 32. The distribution of variance was similar for both estimation methods at the lower levels, but the

---

The proportion of variance explained for each level in the model is shown below in Table 32. The distribution of variance was similar for both estimation methods at the lower levels, but the
MCMC procedure allocated less variance to the facilitator level, and more to the school level. The combined school plus facilitator level was around 16% for both estimation methods, similar to that established for NZ schools generally. The variance explained at the facilitator level was non-significant when estimated using MCMC. As with the previous analyses, this level is retained for consistency, and to allow comparisons of the value-added facilitator effect. The variance of the residual term, which includes differences over time (as well as other, unknown sources of variance), was both large and significant. Student-level variance was also large, with more than a third of the variance in asTTle scores explained at the student level. Year level was a small, but significant source of variance, while the variance at the school and facilitator levels was slightly larger, but of marginal significance. This is due to larger standard error in the variance estimates at the higher levels – indicative of the reduced information available at these levels.

Table 32. Unconditional models comparing variance proportion at each level by estimation procedure.

<table>
<thead>
<tr>
<th></th>
<th>RIGLS</th>
<th>MCMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facilitator</td>
<td>9.3%*</td>
<td>8% (n.s.)</td>
</tr>
<tr>
<td>School</td>
<td>6.3%**</td>
<td>8.1%*</td>
</tr>
<tr>
<td>Year cluster</td>
<td>5.4%***</td>
<td>5.6%***</td>
</tr>
<tr>
<td>Student</td>
<td>38.8%***</td>
<td>35.6%***</td>
</tr>
<tr>
<td>Time (+ residual)</td>
<td>43.2%***</td>
<td>42.7%***</td>
</tr>
<tr>
<td>Deviance</td>
<td>180020.5</td>
<td>173072.8</td>
</tr>
</tbody>
</table>

* p < 0.05; ** p < 0.01; *** p < 0.001

The unconditional model, with only time added as a covariate, is shown in the equations below. These equations indicate that initial achievement among students in this cohort was approximately 52 points lower than the asTTle normative expectations, and that progress was around 38 points higher per year (overall progress SDU = 0.75). The unexplained terms in each equation represent the error, or residual terms, for each level.

\[
ScoreVsExpectation = \beta_{0jklm} + 37.62(0.79)(Time - 1)_{ijklm} + e_{ijklm}
\]  

(20)

Where: \( \beta_{0jklm} = -51.74(8.0) + g_{0m} + f_{0lm} + v_{0klm} + u_{0jklm} \)
The parameter estimates and their associated standard errors for the final model are shown for each of the factors that explained a significant portion of variance in student asTTle scores in Table 33. There were no differences in baseline scores according to the proportion of minority students attending a school, and progress was also equivalent to the overall cohort. Students attending intermediate schools had initial achievement lower than students in other school types, but this difference was variable (reflected by the wide standard error) and non-significant. The factor was retained in the model since a significant portion of variance was explained with its inclusion. Interpretation is confounded by the finding that students in Year 7 had initial achievement much higher than other students. Progress was equivalent by school type, suggesting that this (non-significant) gap is likely to have remained over the duration of the project. However, students in Year 7 had higher progress rates than other students, so the suggested deficit among students attending intermediate schools is unlikely to be meaningful. Older students generally had higher baseline scores, but progress rates were equivalent by year level except among Year 7 students (SDU = .25 more than other students in the cohort).
Table 33. Final HLM showing parameter estimates by estimation method for 2006-2007 writing cohort.

<table>
<thead>
<tr>
<th></th>
<th>RIGLS estimates</th>
<th>MCMC estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unstandardized Coefficients</td>
<td>Standardized Coefficients</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td>Intercept</td>
<td>-131.03</td>
<td>10.37</td>
</tr>
<tr>
<td>Year 5</td>
<td>23.45</td>
<td>5.14</td>
</tr>
<tr>
<td>Year 6</td>
<td>44.67</td>
<td>5.13</td>
</tr>
<tr>
<td>Year 7</td>
<td>59.8</td>
<td>8.23</td>
</tr>
<tr>
<td>Female</td>
<td>50.98</td>
<td>2.58</td>
</tr>
<tr>
<td>Māori</td>
<td>-22.89</td>
<td>3.05</td>
</tr>
<tr>
<td>Pasifika</td>
<td>-16.6</td>
<td>4.45</td>
</tr>
<tr>
<td>Mid Decile</td>
<td>57.23</td>
<td>7.79</td>
</tr>
<tr>
<td>High Decile</td>
<td>64.64</td>
<td>8.67</td>
</tr>
<tr>
<td>Intermediate</td>
<td>-21.77*</td>
<td>19.78</td>
</tr>
<tr>
<td>School Size</td>
<td>.061</td>
<td>.026</td>
</tr>
<tr>
<td>Time</td>
<td>44.93</td>
<td>1.32</td>
</tr>
<tr>
<td>Female * time</td>
<td>-5.11</td>
<td>1.54</td>
</tr>
<tr>
<td>Year 7 * time</td>
<td>12.87</td>
<td>2.38</td>
</tr>
<tr>
<td>Mid Decile * time</td>
<td>-12.98</td>
<td>1.73</td>
</tr>
<tr>
<td>School Size * time</td>
<td>-.028</td>
<td>.005</td>
</tr>
</tbody>
</table>

*This factor is non-significant at baseline, but is of value to the model since model deviance is reduced considerably ($\chi^2 > 75, p < .001$).

Baseline scores by gender and ethnicity typically showed the same differences previously indicated, with a baseline advantage among female students of half a standard deviation, and a disadvantage of around 15 to 20 points for Māori and Pasifika students. Māori and Pasifika progressed at the same rate as the cohort overall, suggesting considerable catch-up against the asTTle normative expectations, but no improvement relative to other students in the project. Female students progressed slower than male students, though this difference was small (SDU = .1). Students attending mid and high decile schools also had baseline scores higher than
students in low decile schools (~60 points). This advantage declined among students in mid
decile schools (relative decline SDU = .25) while students in low and high decile schools had
similar rates of progress. Students in larger schools also had baseline scores marginally higher
than students in smaller schools, with a difference of around 25 points between smaller schools
(~50 students) and larger schools (~450 students). This initial advantage is counteracted by lower
progress rates, with the relative coefficient estimates indicating that achievement was similar for
schools of different sizes by EOY2.

The overall estimate of progress in Table 33 reflects a gain of almost a full standard deviation
relative to the normative sample; this reflects progress when all of the covariates are at zero.
Since progress rates were fairly consistent across subgroups within this cohort, this means that
this is the expected rate of progress among male students of any ethnicity in Year 4, 5 or 6
attending a low or high decile school of any type and of average size with any proportion of Māori
and Pasifika students. Since there was considerable cohesion in the rates of progress in this
cohort, this reflects the gain for many students, but to get an overall picture of the average
project-level progress, the initial model discussed above where time is the only covariate should
be used, since this provides the ‘average’ achievement and progress among all students in the
cohort.

The proportion of remaining variance and variance explained by estimation method is shown
in Table 34, below. The remaining unexplained variance was mostly at the student and time
levels. Less than five per cent of the variance at the student level was explained, while around
one-fifth of the variance in progress was explained. Almost half of the year-level variance was
explained, and almost all of the variance at the school level. None of the variance at the facilitator
level was explained, with most of the remaining variance across schools allocated to the facilitator
level. In contrast to the reading cohorts, this suggests considerable variation in the effect of the
project by facilitator; this possibility is explored further in the value-added section.
The results were again very stable across estimation methods, with similar point estimates for all covariates, and no differences in substantive interpretations. However, the allocation of variance at each level again differed by procedure, though considerably less than for the reading cohorts. Conversely, the results were less comparable to the estimates derived using single level regression in section 4.2.2.3. Several factors were identified as having a significant effect on student outcomes by one procedure and not the other. For example, the single level regression model indicated that being in a school with a large proportion of minority students had a negative effect on progress, while being in schools of different sizes had no effect. The HLM analyses indicated the opposite, with minority proportion having no significant effect, and school size having a modest effect on outcomes. By gender, the results were significant in both instances, but were the reverse of the other. The regression model suggested an additional gain for females (SDU = .22) while the HLM estimated progress among females as slower than among males (SDU = -.1). Additional differences are discussed in Chapter 6, and comparisons are also made to the effect size analyses.

### 5.2.4. Writing Focus – 2008-2009 Cohort

Table 35 shows the distribution of variance by level. The estimates are similar for both estimation methods, but the MCMC procedure yielded lower model deviance. The variance explained at the facilitator level was non-significant irrespective of the estimation procedure, but

Table 34. Remaining variance and variance explained by estimation method.

<table>
<thead>
<tr>
<th></th>
<th>Proportion of remaining variance</th>
<th>Proportion of variance explained</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RIGLS</td>
<td>MCMC</td>
</tr>
<tr>
<td>Facilitator</td>
<td>13.8%**</td>
<td>15.3%*</td>
</tr>
<tr>
<td>School</td>
<td>.9% (n.s.)</td>
<td>.2% (n.s.)</td>
</tr>
<tr>
<td>Year cluster</td>
<td>3.1%***</td>
<td>3.6%***</td>
</tr>
<tr>
<td>Student</td>
<td>41.8%***</td>
<td>41.2%***</td>
</tr>
<tr>
<td>Time (+ residual)</td>
<td>40.4%***</td>
<td>39.8%***</td>
</tr>
<tr>
<td>Reduction in Deviance</td>
<td>3595.4***</td>
<td>4354.8***</td>
</tr>
</tbody>
</table>

* p < 0.05; ** p < 0.01; *** p < 0.001
the level is once again retained for consistency, and to allow comparisons of the value-added facilitator effect in subsequent analysis. The largest proportion of variance was at the 'time' level; unsurprising since this also includes the largest error residuals. Student-level variance was also large, with slightly more than one-third of the variance in students' asTTle scores occurring at the student level. There was more variance at the year level in this cohort, representing approximately one-tenth of the score variance. School level variance was much lower than the usual school-level variance for New Zealand (i.e., ~17%), and much smaller than for previous cohorts; even if the (negligible) variance at the facilitator level is included, the combined proportion of variance at the school plus facilitator level is less than ten per cent.

Table 35. Unconditional models comparing variance proportion at each level by estimation procedure.

<table>
<thead>
<tr>
<th></th>
<th>RIGLS</th>
<th>MCMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facilitator</td>
<td>2.4% (n.s.)</td>
<td>1.0% (n.s.)</td>
</tr>
<tr>
<td>School</td>
<td>6.2%*</td>
<td>7.9%**</td>
</tr>
<tr>
<td>Year cluster</td>
<td>9.6%***</td>
<td>9.7%***</td>
</tr>
<tr>
<td>Student</td>
<td>36.2%***</td>
<td>36%***</td>
</tr>
<tr>
<td>Time (+ residual)</td>
<td>45.5%***</td>
<td>45.3%***</td>
</tr>
<tr>
<td>Deviance</td>
<td>191011.1</td>
<td>183617.4</td>
</tr>
</tbody>
</table>

* p < 0.05; ** p < 0.01; *** p < 0.001

To provide a baseline picture of the overall project effectiveness, the unconditional model is shown below with time as the only covariate. Baseline achievement among students in this cohort was initially much lower than the asTTle normative expectations (93 points below norm, on average), but progress was also much more rapid (overall progress SDU = 1.0). The unexplained terms in each equation represent the error, or residual terms, for each level.

\[
ScoreVsExpectation = \beta_{0,ijklm} + 50.5(0.8)(Time - 1)_{ijklm} + e_{ijklm}
\]

Where: \( \beta_{0,ijklm} = -93.47(7.33) + g_{0m} + f_{0lm} + v_{0klm} + u_{0,ijklm} \)
Parameter estimates and associated standard errors are shown for each factor in Table 36. There were no baseline differences by school size, or proportion of minority students; nor did the progress rates differ according to these factors. There was a large gender gap at baseline in this cohort, with female students having initial scores more than 40 points higher than males. Progress rates were similar by gender, so this gap remained equivalent over the course of the project. Generally older students had higher initial achievement, especially those in Years 6 and 7, though progress rates were marginally lower among Year 7 students. However, students in Year 7 who were attending an intermediate school had lower initial achievement, but greater progress over the course of the project. Effectively the relative magnitudes of the coefficients indicate that by EOY2, scores did not differ markedly by school type.

Initial achievement was higher among mid and high decile schools compared with students attending low decile schools, with a large difference between the typical achievement at low versus high decile schools (SDU = .65). Students in mid decile schools progressed marginally faster than those in low decile schools (SDU = .1), while those in high decile schools progressed more slowly. This decline was not large, but does represent a catch up among students in low decile schools of around one-fifth of a standard deviation relative to students attending higher decile schools. Māori and Pasifika typically had lower initial achievement (around 20 points), and progressed at a similar rate to the overall cohort.
Table 36. Final HLM showing parameter estimates by estimation method for 2008-2009 writing cohort.

<table>
<thead>
<tr>
<th></th>
<th>RIGLS estimates</th>
<th></th>
<th>MCMC estimates</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unstandardized</td>
<td>Standardized</td>
<td>Unstandardized</td>
<td>Standardized</td>
</tr>
<tr>
<td></td>
<td>Coefficients</td>
<td>Coefficients</td>
<td>Coefficients</td>
<td>Coefficients</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td></td>
<td>Beta</td>
<td></td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-171.63</td>
<td>9.44</td>
<td>-169.29</td>
<td>10.66</td>
</tr>
<tr>
<td>Female</td>
<td>44.43</td>
<td>2.28</td>
<td>43.68</td>
<td>2.28</td>
</tr>
<tr>
<td></td>
<td>19.49</td>
<td></td>
<td>19.16</td>
<td></td>
</tr>
<tr>
<td>Year 5</td>
<td>35.95</td>
<td>3.02</td>
<td>36.54</td>
<td>4.74</td>
</tr>
<tr>
<td></td>
<td>11.9</td>
<td></td>
<td>7.71</td>
<td></td>
</tr>
<tr>
<td>Year 6</td>
<td>65.22</td>
<td>3.37</td>
<td>65.83</td>
<td>4.98</td>
</tr>
<tr>
<td></td>
<td>19.35</td>
<td></td>
<td>13.22</td>
<td></td>
</tr>
<tr>
<td>Year 7</td>
<td>91.63</td>
<td>4.83</td>
<td>92.73</td>
<td>6.51</td>
</tr>
<tr>
<td></td>
<td>18.97</td>
<td></td>
<td>14.24</td>
<td></td>
</tr>
<tr>
<td>Māori</td>
<td>-21.49</td>
<td>3.24</td>
<td>-21.62</td>
<td>3.16</td>
</tr>
<tr>
<td></td>
<td>6.63</td>
<td></td>
<td>6.84</td>
<td></td>
</tr>
<tr>
<td>Pasifika</td>
<td>-23.15</td>
<td>3.91</td>
<td>-23.36</td>
<td>3.91</td>
</tr>
<tr>
<td></td>
<td>5.92</td>
<td></td>
<td>5.97</td>
<td></td>
</tr>
<tr>
<td>Mid Decile</td>
<td>28.51</td>
<td>11.79</td>
<td>23.74</td>
<td>11.96</td>
</tr>
<tr>
<td></td>
<td>2.42</td>
<td></td>
<td>1.98</td>
<td></td>
</tr>
<tr>
<td>High Decile</td>
<td>69.42</td>
<td>15.28</td>
<td>64.71</td>
<td>14.88</td>
</tr>
<tr>
<td></td>
<td>4.54</td>
<td></td>
<td>4.35</td>
<td></td>
</tr>
<tr>
<td>Intermediate</td>
<td>-59.02</td>
<td>24.78</td>
<td>-59.34</td>
<td>24.55</td>
</tr>
<tr>
<td></td>
<td>2.38</td>
<td></td>
<td>2.42</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>48.9</td>
<td>1.5</td>
<td>48.71</td>
<td>1.52</td>
</tr>
<tr>
<td></td>
<td>32.6</td>
<td></td>
<td>32.05</td>
<td></td>
</tr>
<tr>
<td>Year 7 * time</td>
<td>-8.56</td>
<td>2.54</td>
<td>-8.32</td>
<td>2.61</td>
</tr>
<tr>
<td></td>
<td>3.37</td>
<td></td>
<td>3.19</td>
<td></td>
</tr>
<tr>
<td>Mid Decile * time</td>
<td>5.29</td>
<td>1.82</td>
<td>5.77</td>
<td>1.87</td>
</tr>
<tr>
<td></td>
<td>2.91</td>
<td></td>
<td>3.09</td>
<td></td>
</tr>
<tr>
<td>High Decile * time</td>
<td>-11.72</td>
<td>2.1</td>
<td>-11.46</td>
<td>2.12</td>
</tr>
<tr>
<td></td>
<td>5.58</td>
<td></td>
<td>5.41</td>
<td></td>
</tr>
<tr>
<td>Full Primary * time</td>
<td>5.18</td>
<td>1.72</td>
<td>5.02</td>
<td>1.78</td>
</tr>
<tr>
<td></td>
<td>3.01</td>
<td></td>
<td>2.82</td>
<td></td>
</tr>
<tr>
<td>Intermediate * time</td>
<td>21.73</td>
<td>3.96</td>
<td>20.73</td>
<td>4.16</td>
</tr>
<tr>
<td></td>
<td>5.49</td>
<td></td>
<td>2.35</td>
<td></td>
</tr>
</tbody>
</table>

For this cohort, the ‘time’ covariate is almost equivalent to the progress indicated for the overall cohort. Within the final model, this reflects progress for each year of the project among students in Years 4 to 6, attending low decile, contributing schools. Conversely, the intercept is much lower than in the baseline model. In this case, the intercept reflects average initial achievement among male students in Year 4, of NZ European or ‘other’ ethnicity, attending low decile contributing or full primary schools. Once again, the initial model discussed earlier in this
section, in which time is the only covariate, provides a more general picture of the average project-level progress for the cohort overall.

An indication of the amount of variance explained, and the distribution of the remaining variance by level, is shown for each estimation method in Table 37. In the current cohort, the RIGLS estimation procedure could not produce plausible estimates when allowing the intercept to vary at the year level clustering once additional factors were included in the model. As a result, the intercept was only allowed to vary at the observation, student, school and facilitator levels, and was fixed across year level clusters for the RIGLS estimates. Therefore, the RIGLS model could not explain variance at year level; indicated by ‘N/A’ in the table. The MCMC estimation is more robust and produced plausible estimates but, due to the nature of MCMC, it is not possible to be certain whether the model has ‘converged’ since the method (at a very simple level) involves constant sampling, updating the current ‘knowledge’, and then resampling. However, the estimates obtained for the coefficients above were similar for both methods, and the MCMC diagnostics indicated a strong likelihood that sufficient sampling had been completed. The RIGLS estimation indicated no explanation of variance at the facilitator level, while the MCMC procedure estimated a quarter of the variance at this level had been explained. The methods differ at the school level also, though this is simply because the RIGLS procedure allocated some remaining variance to the facilitator level, while the MCMC estimation suggested the remaining variance was almost completely at the school level. Almost one third of the variance at the ‘time’ level was explained.
<table>
<thead>
<tr>
<th></th>
<th>Proportion of remaining variance</th>
<th>Proportion of variance explained</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RIGLS</td>
<td>MCMC</td>
</tr>
<tr>
<td>Facilitator</td>
<td>2.1%</td>
<td>0.1%</td>
</tr>
<tr>
<td>School</td>
<td>7.4%***</td>
<td>8.1%***</td>
</tr>
<tr>
<td>Year cluster</td>
<td>0%</td>
<td>2.5%*</td>
</tr>
<tr>
<td>Student</td>
<td>51.9%***</td>
<td>50.1%***</td>
</tr>
<tr>
<td>Time (+ residual)</td>
<td>38.6%***</td>
<td>38.4%***</td>
</tr>
<tr>
<td>Reduction in Deviance</td>
<td></td>
<td>4957.8***</td>
</tr>
</tbody>
</table>

* p < 0.05; ** p < 0.01; *** p < 0.001

The results were again stable across estimation methods, with similar point estimates for all covariates, and no differences in substantive interpretations. However, the allocation of variance at each level again differed by procedure, though this is confounded by the failure of RIGLS to produce estimates when intercepts were allowed to vary at the year level. This also complicates comparisons of variance explained. The results were also much less comparable to the estimates derived using single level regression in section 4.2.2.4. Several factors were identified as having a significant effect on student outcomes by one procedure and not the other. For example, the single level regression model indicated that Māori and Pasifika students gained less than the overall cohort, while the HLM analyses reported no differences by ethnicity. In addition, the single level regression model found no advantage of being in low or mid decile schools relative to those in high decile schools, while the HLM analyses indicated the opposite, with students in high decile schools progressing more slowly than students in low and mid decile schools (SDU = -.22 to -.32). Remaining differences are discussed in Chapter 6, alongside comparisons to the effect size analyses.

### 5.2.5. Value-added at the facilitator level

This section uses the final hierarchical linear models developed in the previous sections to examine whether there were differences in the degree of value-added by the project depending on the facilitator working with the set of schools. This was assessed by comparing the average
residual scores of student outcomes depending on the facilitator who worked with their school. This is essentially the same process as for the value-added models based on conventional regression, and is based on the same premise; that differences in student progress by facilitator can be compared, with the understanding that any differences are not simply a result of the factors included in the models (or explicitly excluded due to lack of variance explained). This procedure is similar to that used by Kane and Staiger (2008) in their evaluation of teacher effects on student achievement outcomes, and is a form of the nested random effects model discussed by Rose, Henry and Lauen (2012). Since HLM partitions the variance by level, this is a more accurate representation of the estimated added value by facilitator because the model attempts to specifically exclude variation in student outcomes from other levels. Therefore, if the student residual scores at the facilitator level differ markedly from zero (i.e., the average project effect), one can assume that the facilitator involved contributed more or less (depending on whether the residual is positive or negative) than other facilitators. As with the conventional regression section, in order to be as confident as possible that any differences related to the facilitator rather than the school, facilitators’ residual scores from both cohorts were combined and only facilitators with data for more than 100 students and more than three schools were included in the analysis. To allow direct comparison with the value-added section of the previous chapter, the analyses remain separated by focus, but aggregated over time. The MCMC estimates are used due to the increased performance with non-normal and missing observations, and when working with three of more levels, as explained in the introductory section.

5.2.5.1. Reading Focus – both cohorts

Among schools focusing on reading, there were 17 facilitators with data from more than three schools and at least 100 students, as indicated in the previous chapter. However, in contrast with the estimates of facilitator effect derived from conventional regression, the HLM residuals indicate that there were no significant differences by facilitator, as shown in Figure 10. This figure indicates the estimated prediction error for each facilitator, after taking into account the factors
included in the model, and in all cases the confidence intervals overlap zero. Zero is defined as
the average project effect, indicating that there was no difference in the gains among students
attending schools working with different facilitators. This is in stark contrast to the value-added
findings for reading in Chapter 4, in which the facilitators identified as having the highest gains
had an estimated added value around 40 points greater than those with the lowest gains.

Figure 10. Differences in average value added by facilitators for reading.

5.2.5.2. Writing Focus – both cohorts

As indicated in Chapter 4, there were 18 facilitators with data from more than three schools
and more than 100 students. In contrast with the reading cohorts, there was some variation in
outcomes by facilitator in the writing cohorts. The results for students attending schools working
with three facilitators were significantly (though only marginally so) higher than estimated after
controlling for the variation related to the factors included in the models. This reflected an
estimated advantage of 16-18 points for students in the schools working with these facilitators,
relative to the project overall. Conversely, there were four facilitators for whom results were
significantly (again only marginally) lower than predicted, reflecting an estimated average disadvantage of 14 to 23 points relative to the overall project. Since the overall expected project gain was equivalent to almost one hundred asTTle points greater than the progress estimated for maturational effects alone, this indicates that all facilitators provided considerable added value, though some were even more successful with the schools they were working with. However, these results are again in stark contrast with those presented in Chapter 4. Although differences among facilitators remain, the facilitators identified as being differentially effective compared with the project overall, are different from those identified in Chapter 4. Indeed, one facilitator that was identified as having added significantly less value using the single level regression models, now has an estimated residual gain of approximately 20 points – the highest for all facilitators.
5.3. Chapter Summary

The hierarchical linear models developed in this chapter incorporate as much of the data context as possible, producing the most nuanced form of analysis presented in this thesis. These analyses also allow achievement and progress to be explicitly estimated within the same models, assisting with clearer conclusions about project effectiveness and remaining differences. The general models indicated modest overall gains among students in the reading cohorts relative to the asTTle normative sample, with accelerated progress around 1.4 times greater in the 2006-2007 (additional SDU = .2) and around 1.6 times faster in the 2008-2009 cohort (SDU = .33), compared with usual maturational progress. Gains were much larger in writing, with the HLM estimates indicating that progress was 2.8 times faster than usual in the 2006-2007 cohort (SDU = .75) and 3.5 times faster in the 2008-2009 cohort (SDU = 1.0). The absolute magnitude of

![Figure 11. Differences in average value added by facilitators for writing.](image)
these estimates differed, sometimes considerably, from those reported in Chapters 2 and 3, but all analyses indicated larger effects for writing than for reading, and higher gains in the 2008-2009 cohort than in the 2006-2007 one.

Conclusions were similar regardless of the estimation procedure, with some notable exceptions. In all cases, the proportion of variance allocated to the facilitator level in the unconditional models was smaller for MCMC estimation compared with the RIGLS estimates. In addition, while the estimated coefficients were generally similar for both procedures, the degree of uncertainty sometimes varied markedly, with considerable variation in the magnitude of the standard error estimates for the same factors. For example, Pasifika students attending schools with a reading focus in the 2008-2009 cohort were estimated to have baseline scores approximately 29 points lower than the reference category (NZ European and ‘other’ ethnicities) by both procedures. However, the estimated standard error produced by the MCMC procedure was half that produced using RIGLS, resulting in a large difference in the standardised beta. Since the main focus is on the unstandardized coefficients, the difference has minimal impact but, in other cases, where one procedure would exclude the factor from the model, the other would include it, due to the differences in standard error. In the 2006-2007 reading cohort, there were two covariates that would have been excluded by the MCMC procedure that were included by RIGLS; specifically, the baseline differences for Year 7 students, and for those attending schools with differing proportions of minority students. There were also large differences in the estimates of variance explained at the higher levels depending on the procedure used. This is unsurprising; since there is less information available at higher levels, the stability of the estimates decreases. As noted previously, the MCMC procedure typically performs better than RIGLS when there are three or more levels, especially with respect to interval estimates (i.e., the standard error) (Browne & Draper, 2006), so it is likely that the MCMC estimates provide a better fit for the data.

The results of the hierarchical models generally indicate that the progress rates were remarkably consistent across subgroups. Progress rates were greater for older students in all cohorts. In the 2008-2009 writing cohort, although the progress for Year 7 students is marginally lower, the progress rate for students in intermediate schools was much higher. Considered in
combination, this resulted in the estimates of progress among Year 7 students being marginally larger than students in other year levels. This finding was more prominent in the reading cohorts, which is likely to be an artefact of the expected progress rates for asTTle, which decline from an expected gain in reading of 50 points in Year 4 to nine points in Year 8. Although there were some differences in the progress rates among students attending schools of different types, sizes and proportion of minority students, the actual magnitude of the effects is generally small.

In addition, although there were differences in progress by decile group, the HLM estimates indicated that these were also typically not large; in contrast with the findings from previous chapters. In reading schools, students attending higher decile schools typically gained more during the project, but in the 2006-2007 cohort this was mitigated by the finding that students attending schools with higher proportions of minority students also progressed more rapidly. The average proportion of NZ Māori and Pasifika students attending low decile schools in New Zealand is 72% compared with 12% attending high decile schools; a difference of 60% (Education Counts, n.d.). As a result, the expected gain for students in the 2006-2007 reading cohort with the average proportion of minority students for the relevant decile band is actually equivalent, irrespective of decile. There was, however, a moderate increase in the decile gap in the 2008-2009 reading cohort. Conversely, in the writing cohorts the decile gap either stayed approximately equal, or narrowed.

The value-added analyses indicated no differences in the gains of students attending schools working with different facilitators in the reading cohorts, despite sometimes large differences in the progress rate of students attending different schools. These results contradicted the value added models presented in Chapter 4, in which differences were found by facilitator. For schools with a writing focus, significant differences in the apparent value-added by facilitators were indicated. The gain among students working with the facilitator whose estimated value-added was highest typically had gains 40 points greater than those working with the facilitator estimated to have added the least value. As noted previously, on average, all facilitators elicited considerable gains in writing, with the average gain among students working with the least effective facilitator still greater than 70 asTTle points. The average yearly gain indicated by the
normative sample for asTTle writing is 20 points, so this remains a large effect. However, while
the regression analyses also indicated differences in the facilitator effect, the rank order of
facilitators differed considerably, with one of the least effective facilitators identified by single level
regression appearing to be the most effective using HLM.
CHAPTER 6. DISCUSSION

This chapter summarises and compares the conclusions made about student progress during the Literacy Professional Development Project, based on the analyses presented in each chapter. Initial discussion focuses on the overarching research question of the thesis, identifying the differences in the results made about the project gains according to the analysis methodology to determine the extent to which these differences would have substantive effects on conclusions. The secondary aim, to investigate project effectiveness, is addressed in the following section (6.2) with respect to the additional research questions. Specifically, this section considers whether the professional development of teachers transferred to a measurable shift in student outcomes, and, whether these differences resulted in reductions in between-group differences. Section 6.3 offers a concluding statement summarising the findings and contribution of the thesis.

6.1. Does choice of analysis have a substantive effect on conclusions made about the results when analysing a large educational dataset?

The primary aim of this thesis was to assess the extent to which choosing a particular method of analysis over another had implications for conclusions. Several commonly used quantitative methodologies were employed, and the results are compared in this section to identify similarities and differences. The LPDP dataset was chosen since this was a large, naturalistic dataset with four different cohorts of students (the two foci, plus the two cohort intakes), allowing consideration of the relative performance of the chosen methods across comparable contexts. A large dataset was important, since it is frequently asserted that various assumptions of the analysis are relatively unimportant when enough data (defined by different authors as anywhere between 30 and several hundred) have been collected (i.e., Altman & Bland, 1995; Elliott & Woodward, 2007; Ghasemi & Zahediasl, 2012; Pallant, 2007). The use of datasets from actual schools, with more than a thousand students in each dataset, allows this assertion to be tested with respect to how
reliance on the robustness of methodology might have implications for conclusions and, potentially, on policy and students as a result of these conclusions.

Five main statistical methodologies were used to estimate the project effect for different subgroups throughout this thesis: the standardised mean difference (SMD); the non-parametric equivalent, Cliff’s delta, transformed to standard deviation units (NPSMD); the estimates taking regression to the mean into account explicitly (RTM); the single level, or convention regression models (SLR); and, the hierarchical linear models (HLM). Since the point estimates obtained using the MCMC estimation were virtually identical to those obtained using RIGLS, the estimates from each are not considered separately here. The extent to which these methods resulted in different estimates of progress for each of the main subgroups, compared with the progress estimated by each method for the relevant cohort, is shown in Figure 12, below. The zero-line reflects progress equivalent to the rest of the cohort, so there is no reason to expect estimates to overlap this point. Indeed, instances where the lines extend across the zero line indicate that, for at least one estimation method, average gains for the subgroup were significantly larger than other students in the cohort, while at least one estimation method indicated gains were significantly smaller than other students. In other words, in these cases, not only did the methods differ with respect to absolute magnitude of the estimated effects, they also disagreed about the direction of difference. Each line represents one of the four different datasets, and the discrepancies between the estimation methods are indicated by the length of the lines for each covariate.
Figure 12. Difference between the estimated progress for each subgroup from the overall cohort gains (the zero line), showing the range of estimates across analysis type.

The smallest variation by method is shown for the male, female and NZ European subgroups, which is likely to at least partly reflect the larger sample sizes of these subgroups. Despite the lower variation in the estimates derived by the five different methods for these groups, there were still some substantive differences that arose. For example, there was a difference of .15 standard deviations between the lowest and highest estimates of progress for males in the 2006-2007 writing cohort. At one extreme, progress is estimated to be slightly faster than the overall cohort (HLM estimated SDU = .05), while another method estimated that progress was slower than average for the cohort (SLR estimated SDU = -.11). Even though these differences were the smallest, the differences in directionality are concerning, and would likely to result in substantive differences in the conclusions made about the progress of these subgroups.

The variation in the estimated progress relative to the overall cohort was much larger by decile and for minority ethnicities, with differences of up to one-quarter of a standard deviation. The progress of Pasifika students in the 2006-2007 reading cohort was estimated to be significantly faster than for the overall cohort by the SMD (SDU = .16), and significantly slower by
the RTM procedure (SDU = - .11), for example. This is an overall difference of more than a quarter of a standard deviation – approximately equal to the learning expected in a typical schooling year.

There were also considerable differences in the absolute estimates of progress for each subgroup depending on the method used. The following four figures show the estimated progress in standard deviation units for each subgroup, based on the five procedures. Recall that usual maturational effects have been controlled for, so the zero line in these graphs reflects progress equal to usual maturation. Within each individual factor, the width of the variation shows the degree to which the different procedures provided different indications of progress. The dashed lines have been added so that the reader can identify easily which of the procedures tended to yield higher, or lower estimates compared with the other procedures. In these four figures, “overall” is also shown as a factor, indicating the range of estimated project gains for each cohort. Note that the estimate of overall gain is always the same for SMD and SLR. This is because both procedures calculate the overall average gain in an equivalent way. Differences in the subgroup estimates obtained by these two procedures arise due to the inclusion of the other factors into the regression model, while the SMD estimates are simply pairwise comparisons.

For the 2006-2007 reading cohort, shown in Figure 13, the estimated gain for students within the project relative to those in the normative sample varied from 16 to 34 asTTle points (SDU = .16 to .34). The NPSMD provided the most conservative estimate of gain, while the largest estimated gain was obtained using the procedure to account for RTM. Typically the RTM estimates in this cohort were higher than those derived using other procedures, with the notable exceptions of the estimates for NZ Māori and Pasifika. Since the RTM boosts estimated gains for those achieving above the mean, and lessens estimates for those below the mean, this is unsurprising. The most conservative estimates were generally obtained using NPSMD and HLM, both of which take non-normality into account. The most variable estimates appear to have been obtained using SLR. While the SLR estimates were generally about average, it was the only procedure that produced estimates at both extremes, producing the lowest estimates of gain for NZ Māori students, and for those in low decile schools and the highest estimate of progress for
students in high decile schools. This suggests that if there is bias in the estimates obtained using SLR, the bias is not directionally bounded. Generally the estimates followed the same direction; that is, all procedures indicated that progress was lower than the overall cohort for students in low decile schools. However, while HLM, SLR and RTM estimated that gains were higher among students in high decile schools, the NPSMD and SMD estimated that gains were lower. This is likely to be a result of the failure of both effect size procedures to take baseline score into account.

Figure 13. Estimated gains made by subgroup in the 2006-2007 reading cohort by procedure.

The estimated gain for students in the 2008-2009 reading cohort, relative to those in the normative sample varied from 28 to 42 asTTle points (SDU = .28 to .42). The most conservative estimates are again typically obtained using the NPSMD, while the highest estimates were generally obtained using SMD and SLR. The RTM estimates showed the greatest variability, predicting higher progress among NZ European and high decile students, and lower among NZ Māori and Pasifika, compared with the other procedures, which again reflects the nature of this procedure. The HLM estimates were generally about average, except for the indicated progress among students attending low decile schools, for whom the HLM procedure suggested progress of fewer than 20 asTTle points (SDU = .17). Again, the SMD procedure produced estimates that occasionally differed from the other procedures in terms of the direction from the overall cohort.
Progress among students attending low decile schools was estimated to be lower by the SLR, HLM and RTM procedures, but higher by the SMD (and negligibly so for the NPSMD).

Figure 14. Estimated gains made by subgroup in the 2008-2009 reading cohort by procedure.

Overall gains in writing were larger than for reading, irrespective of the procedure used to estimate these gains (see Figure 15 and Figure 16). However, there were large differences in these estimates, with gains of around half a standard deviation suggested by the NPSMD and RTM for the 2006-2007 writing cohort, while the indicated gain using HLM was almost 50 per cent larger (SDU = .75). As with the reading cohorts, the NPSMD generally produced conservative estimates. In contrast with the reading cohorts, however, the HLM procedure generally indicated that greater gains were made by all subgroups than the gains indicated by other procedures. Since this cohort had low baseline scores, the RTM estimates are generally lower than for the other procedures. The results produced by the SMD were typically about average. In this cohort, there were some interesting differences in the directionality of the progress estimates. For example, progress was indicated to be higher among male students by the HLM, the SMD and the NPSMD, but lower by the RTM and SLR.
The estimated gains made by students in the 2008-2009 writing cohort were even more variable than any of the previous cohorts, with estimated gains relative to the normative sample ranging from 0.61SDU to 1.01SDU, a difference of around 40 asTTle points (see Figure 16). The average expected gain in asTTle is 20 points per year, so this difference reflects a difference of two years of business-as-usual schooling, simply due to the procedure chosen to estimate the gains. The RTM estimates were clearly the most conservative, indicative of how far below the mean students in this cohort were at baseline. The highest estimated gains were obtained using HLM, which also indicated that progress was similar for all subgroups except for those in mid (estimated progress slightly higher) or high (estimated progress slightly lower) decile schools. The estimates derived using the NPSMD, the SMD, and the SLR procedures produced the most ‘average’ results. There were no marked differences in the direction of the estimates for this cohort, except for a slight gain for students in low decile schools indicated by HLM, compared with a slight decline indicated by SMD and NPSMD, and a larger decline by the SLR and RTM procedures.
There were also considerable differences in the results of the value-added analyses based on the SLR versus the HLM models. In reading, SLR suggested that there was a significant facilitator effect, while the HLM analyses indicated that any differences were extremely small and non-significant. Conversely, in writing, the SLR analyses indicated that there were five facilitators who elicited significantly greater gains than the cohort average, while the gains among students attending schools working with three facilitators were significantly lower than average. In contrast, the HLM analyses indicated that there were only three facilitators who apparently added more than the average value-added by the project, and four who added less. Some variation is unsurprising since the HLM analyses explicitly partition variance at the facilitator level from variance at other levels, which SLR cannot do. However, the concerning aspect of these differences is that the facilitators identified as adding greater or lesser value by the two methods did not correspond with each other. Indeed, the facilitator identified as having added the second least by the SLR, was estimated to have added the most value by the HLM.

*Figure 16. Estimated gains made by subgroup in the 2008-2009 writing cohort by procedure.*
6.2. Additional research questions

The secondary aim of the thesis was to evaluate the project effectiveness of LPDP. Specifically, the research aimed to investigate whether the professional development provided during LPDP transferred to gains in student achievement. This is an important question since many previous professional development interventions fail to address the impact of the professional development on student achievement, instead focusing solely on change measured at the teacher-level (Timperley & Alton-Lee, 2008). Along with overall project gains, an additional research question was whether there were reductions in between-group differences as a result of the project, to determine whether gains were equitable across demographic profiles. Since this thesis uses non-simulated data, it is extremely difficult to make assertions about which estimate is the most accurate, or unbiased. Rather, the intent is to demonstrate how these differences could have real and meaningful implications for students and educational policy. It was intended that the use of real data would have greater significance for applied researchers in education. However, as a result of this decision, it is not possible to answer the question about the success of the project by using just one of the procedures presented throughout the thesis – particularly in light of the large differences produced. Instead, all estimation methods are considered, and the average SDU used to indicate the degree of success.

6.2.1. LPDP success at improving reading achievement and reducing group differences

The success of LPDP in producing shifts in the 2006-2007 reading was small (average estimated SDU = .23). At risk students (i.e., male, low decile, NZ Māori and Pasifika) typically gained less from this cohort. However, no subgroup made less progress than the normative sample, irrespective of the estimation procedure. The degree of shift was higher in the 2008-2009 cohort (average SDU = .37). All subgroups progressed more quickly than the normative sample, though all procedures estimated that students in high decile schools gained more, increasing the socio-economic gap. All procedures suggested that gains were generally similar by ethnicity however, which is positive. The results suggest that the project was moderately successful in
reading, with an acceleration of around 12 to 18 months of schooling during the two years. Students in priority subgroups also typically experienced accelerated shifts, but sometimes these were less than for the overall cohort, indicating that the professional development had minimal effect on reducing specific disparities among the subgroups within the reading cohorts.

### 6.2.2. LPDP success at improving writing achievement and reducing group differences

The gains made by students in both writing cohorts were large, irrespective of the procedure used to estimate the gain. Again the project’s effectiveness appears to have improved over time, with larger gains in the 2008-2009 cohort (average SDU = .82) relative to the 2006-2007 cohort (average SDU = .61). The usual gain in asTTle writing is 20 points per year, so this reflects gains of three to four years more than the normative sample, during the two year project. Students in low decile schools in the writing cohorts progressed at or above the rate of the overall cohort, as did Pasifika in the 2006-2007 cohort. However, NZ Māori progressed marginally more slowly than students of other ethnicities in this cohort, while both Māori and Pasifika made slightly lower gains than the rest of the 2008-2009 cohort (average SDU = .75). All subgroups gained at least half a standard deviation compared with the normative sample, for both cohorts. The results suggest that the project was very successful in writing, with progress among students in the project more than two to three times that which would usually be expected. It also appears that the project became more successful over time, since the average estimated effects were larger in the latter cohorts for both reading and writing. However, despite the very positive overall shifts, the gains among priority subgroups were similar to, or sometimes less than for the overall cohort. This indicates that the professional development had minimal effect on reducing specific disparities among the subgroups within the writing cohorts.
6.3. Concluding statement

This thesis made use of four large, real-world datasets to investigate the effect of analysis choice on the conclusions that would be drawn about the data. These datasets were comparable to much of the data in the educational arena, with some degree of non-normality and attrition to be expected in real-world data. Mitigating features of these datasets were the number of observations, which was comparatively large, and the unit of measurement, which was a continuous scale that was developed using item response theory. However, despite these mitigating factors, the results of this thesis demonstrate that researchers’ decisions about which particular analysis to use can potentially produce large and substantive differences in conclusions made about the results. The analyses were based on real data that many researchers will relate to, so it is hoped that these results will be of particular relevance to applied researchers. Relative gains estimated by the different procedures presented throughout the thesis differed by up to half a standard deviation, which represents a massive difference occurring as a result of the particular features of the analysis (i.e., in terms of what is accounted for and what is not), rather than due to differences that are real and educationally meaningful. In addition, facilitators identified as less effective using one method were identified as more effective by another. This has particular implications for measures of teacher effectiveness, which are based on the same methods. In some jurisdictions, salary is at least partially a function of value-added analyses (NCTQ, 2012), so this finding highlights that it is imperative that such analyses be treated with caution and undertaken with the utmost of care. Other researchers have raised concerns about such measures (i.e., Amrein-Beardsley, 2008; Gorard, 2010; Scherrer, 2011), and the results presented here certainly support these concerns.

It is clear that the professional development provided by LPDP transferred to student gains, particularly in writing. As a result, it is likely that much can be learnt from the methods and practices employed, particularly in light of other findings that the majority of professional development interventions have had limited impact on student achievement and progress. The larger effects in the 2008-2009 cohort suggest that improvements were made in the professional development implementation, so investigation of these improvements could prove useful for
researchers interested in improving outcomes via professional development. The success of LPDP has previously been credited to the development of educational partnerships through coherence within and between the multiple levels of the schooling and administration systems, alongside a focus on evidence-informed inquiry into effectiveness at each level of the system (Timperley & Parr, 2009). The results detailed in this thesis lend credence to the effectiveness of the project model; though further work investigating how to target the transfer of teacher-level professional learning toward more specific subgroups more effectively would be useful.
REFERENCES


http://www.gvglass.info/papers/meta25.html


MacInnes, J. (2009). Proposals to support and improve the teaching of quantitative research methods at undergraduate level in the UK. Economic and Social Research Council.


O'Connell, P. J. (2010). *Is sustainability of schooling improvement an article of faith, or can it be deliberately crafted?* Unpublished doctoral dissertation, ResearchSpace@Auckland.


Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology, 9*, 165–181.


resources. In Resource-adaptive cognitive processes (pp. 95-117). Springer Berlin Heidelberg.


