BUILDING ROBUST STATISTICAL MODELS FOR CATEGORICAL OUTCOME VARIABLES IN EDUCATIONAL RESEARCH

Tong Zhu

Curriculum and Pedagogy

Faculty of Education and Social Works

University of Auckland

This dissertation is submitted for the degree of Doctor of Philosophy

February 2020

To my mom

Zhu Huixiang

ABSTRACT

This thesis had a primary aim: to build robust statistical models with categorical outcome variables for intervention evaluation in educational research. It is critical for policy and equity reasons to know just how effective a school intervention is and under what conditions. Addressing this aim meant solving two challenges in value-added models in educational intervention research in New Zealand: (a) the lack of uniform valued student outcome and (b) the complexity of the data structure.

The first practical challenge leads to the first research question of this thesis: how to choose between valued student outcomes for value-added models to measure the effectiveness of school interventions? Specifically, how to optimise reliability when fitting categorical outcome variables in a continuous outcome model and how to reliably use dichotomised continuous outcome variables in a binary outcome model?

The second practical challenge leads to the second research question of this thesis: how to specify the model to achieve a desirable balance in the trade-off between model complexity and model validity? Specifically, how many underlying random effects should be fitted in a multilevel model with categorical outcomes in order to reliably and effectively determine individual- and group-specific effects.

Answers to these questions will be illustrated by the use of two case studies: one from a sustainability study based in two clusters of primary schools; and another from a large-scale, longitudinal research and development programme based in secondary schools.

The findings from the primary school case study showed that with respect to the first research question: (a) categorical outcome variables can be reliably fitted in a continuous model as long as the underlying distribution is asymptotically normal (such as stanine scores in standardised tests) and (b) dichotomisation of continuous outcome measures results in a more conservative but more direct estimation of school intervention effects.

The findings from the secondary school case study showed that with respect to the second research question: (a) model reliability and accuracy are improved when additional random effects are fitted and (b) the size of model improvement (in terms of reliability and accuracy) varies between random effects where improvements are mostly contributed by the inclusion of the lower-level random effects.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the help and support of the kind people who have kept having faith in me and showed their constant encouragements.

First and foremost, I would like to express my deepest gratitude to my supervisors: Associate Professor Mei Kuin Lai and Professor Stuart McNaughton. Mei has generously given me tremendous support over the past five years. Her guidance and enthusiasm have given me an incentive to follow my dream of being an academic. Stuart has taught me, through his wisdom, knowledge and encouragement, how research is done. I am grateful for his contributions of time and ideas to my thesis.

The members of the Woolf Fisher Research Centre have contributed immensely to my personal and professional time at the Faculty of Education and Social Works. Special thanks to Angela McNicholl, Dr Rebecca Jesson, Dr Aaron Wilson, Victoria Cockle, Naomi Rosedale, Jacinta Oldehaver, Maria Meredith, Selena Meiklejohn-Whiu, Cynthia Orr, Yu Liu, and Rachel Willamson for their friendships, advice, and collaboration.

I would also like to thank Associate Professor Ilze Ziedins and Associate Professor David Scott for inspiring my interests in statistics and perpetuating it throughout the years of my post-graduate study. Ilze has taught me by example the importance of being rigorous and showed me the way of mathematical thinking. David has guided me to step into the world of computational statistics. It is these mentorships that have prepared me to pursue a career in academic research.

My appreciation also extends to my dearest friends. To my dearest, Dr Ekaterina Vinkovskaya, thank you for being there, you are my person. To Ju Li, Yu Yue, Dr Lisa Yizheng Chen, Selena Hsiao, Yvonne Lam, Christine Dong, Dr Jing Liu, Huang Mei, and Peng Yi, thank you for your involvements and your practical and personal support. I am indebted to you for your efforts in helping me succeed.

Finally, my love to my family, to my mother, Zhu Huixiang, to whom I owe a great deal of my accomplishments.

Special thanks to Kayla Friedman and Malcolm Morgan of the Centre for Sustainable Development, University of Cambridge, the UK for producing the Microsoft Word thesis template used to produce this document.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

1 INTRODUCTION

There has been a longstanding national and international focus on improving the achievement of students for policy and equity reasons, and it is critical to know just how effective a school intervention is and under what conditions (Lai, McNaughton, & Zhu, 2018; Lai, Wilson, McNaughton, & Hsiao, 2014; Jesson, McNaughton, Wilson, Zhu, & Cockle, 2018b; Jesson, McNaughton, Rosedale, Zhu, & Cockle, 2018a; Finch & Cassady, 2014; Leckie, 2009; Timmermans, Snijders, & Bosker, 2013). In New Zealand, Māori and Pasifika students have the lowest rates of attainments in the National Certificate of Educational Achievement (NCEA), and this has led to specific government targets around improving Māori and Pasifika students' certificate attainment (Ministry of Education, 2012). These students tend to have high school graduation rates that are significantly lower than other students nationally and score lower in a range of international and local assessments at primary school (Lai et al., 2018; Lai et al., 2014; Jesson et al., 2018b; Jesson et al., 2018a). Factors that may influence students' level of achievement and improvement are well documented and range from students' characteristics including gender and initial achievement level; community factors such as parental guidance and socioeconomic status; and school characteristics such as school culture, teacher characteristics, curriculum, pedagogical content knowledge, resources, and academic counselling supports (Hattie, 2009; Armitage, 2008; Meyer, McClure, Walkey, Weir, & McKenzie, 2009; Lai et al., 2018; Lai et al., 2014; Jesson et al., 2018b; Jesson et al., 2018a).

Several local and national interventions, professional research, and development studies have been carried out specifically to improve Māori and Pasifika students' educational

achievements (Lai et al., 2018; Lai et al., 2014; Jesson et al., 2018b; Jesson et al., 2018a, Bishop, Berryman, & Richardson, 2003; Ministry of Education, 2013; Tuuta, Bradnam, Hynds, & Higgins, 2004), with varying reported levels of intervention success.

There are two challenges in measuring improvements in educational achievement (e.g. high school graduate rate) in New Zealand primary and secondary school systems: one is the lack of uniform valued student outcome measures and the other is the development of rigorous statistical models to determine the added-value that interventions make (Aitkin & Longford, 1986; Braun, 2005; Briggs, 2012; Gelman, 2006, Hattie, 2009; Armitage, 2008; Meyer, McClure, Walkey, Weir, & McKenzie, 2009; Lai et al., 2018; Lai et al., 2014; Jesson et al., 2018b; Jesson et al., 2018a).

This thesis aims to discuss rigorous quantitative research methods to evaluate large-scale longitudinal and cross-sectional intervention studies for improving primary and secondary education achievement in New Zealand. A reliable and efficient quantitative research method for intervention evaluation requires a stable and relevant valued student outcome and a robust statistical model with adequate model reliability and model accuracy.

In the following sections, I first discuss the assessments used to measure student learning outcomes in New Zealand primary and secondary educational systems in Section 1.1, followed by a general review of value-added models in educational research and the application of multilevel models in value-added models in Section 1.2. In Section 1.3, I discuss the aims of this thesis in detail. In Section 1.4 I outline organisations of the remaining thesis.

1.1 Valued student outcomes in New Zealand education system

In New Zealand, students begin their primary education at the age of 5 (Year 1) and continue through to the age of 12 (Year 8). Year 7 and 8 are offered at either a primary school or an intermediate school. Secondary education follows at Year 9 (age of 13) and continues through to Year 13 (age of 18).

Students study subjects under the framework of the New Zealand National Curriculum (NZC) in primary and intermediate schools. The range of subjects includes English, the arts, health and physical education, languages, mathematics and statistics, science, social sciences, and technology (Ministry of Education, 2007). Students at secondary schools study towards the National Certificate of Educational Achievement (NCEA). Some

schools (mostly independent schools) also offer Cambridge International Examinations and International Baccalaureate programmes.

Up to 2017, students in primary and intermediate schools are regularly assessed against expectations for their age level in reading, writing, and mathematics under the guidance of the National Standards (Ministry of Education, 2007). At the time of writing, the National Standards and Ngā Whanaketanga Rumaki Māori have been abolished, and there is at the moment no replacement for these as the Ministry of Education is working in collaboration with the schools, students, parents, whānau, iwi and communities to develop a new framework for progress and achievement across the National Curriculum. For the analyses in this thesis, I will be using data collected under the (abolished) National Standards framework. However, these analyses are generic and able to be applied to any new framework.

Schools in New Zealand have the agency and flexibility to decide how they assess, and what assessment data they collect, report, and analyse at primary schools (Ministry of Education, 2007). In secondary schools, assessments are regulated under the framework of the National Certificate of Educational Achievement.

In subsection 1.1.1, I discuss common assessment tools used in New Zealand primary schools, as well as the New Zealand Curriculum. I then discuss the secondary assessments under the National Certificate of Educational Achievement (NCEA), the main national assessment at secondary schools in subsection 1.1.2.

1.1.1 Valued student outcomes in primary schools

There are diverse measures of learning outcomes across schools (and even within schools across year levels and learning areas) in the primary education system in New Zealand because the New Zealand government sets no mandated assessments. New Zealand schools are self-governing (Ministry of Education, 2007) and primary schools have the agency to choose at their discretion which assessment tools will be used to evaluate student learning outcomes, provided the tools are validly and reliably aligned with the New Zealand Curriculum (Ministry of Education, 2007), as one of the mandatory requirements from the Ministry of Education.

1.1.1.1 The New Zealand Curriculum

The New Zealand Curriculum is an official policy that regulates teaching and learning in English medium New Zealand schools. It aims to set the direction for student learning

and to provide guidance for schools in relation to the taught curriculum. Te Marautanga o Aotearoa serves the same function for Māori-medium schools (Ministry of Education, 2007).

Eight learning areas include English, the arts, health and physical education, learning languages, mathematics and statistics, science, social sciences, and technology are specified by the New Zealand Curriculum (Ministry of Education, 2007). Each learning area has a specific structure and learning objectives. For example, English is structured around two interconnected strands, each encompassing the oral, written and visual forms of the language. The achievement objectives within each strand suggest progressions through which most students move as they become more effective oral, written, and visual communicators.

Not all learning areas across year levels require formal assessment, and excessive assessing is also discouraged by the New Zealand Curriculum (Ministry of Education, 2007). Nonetheless, it sets out achievement objectives for each year level within some learning areas. Those achievement objectives serve as the foundation of alignments between formal assessment tools and the New Zealand Curriculum.

In primary schools, students' abilities in reading, writing, and mathematics are formally assessed against expectations for their year levels, as set out by the National Standards. Take reading for example, according to the National Standards (Ministry of Education, 2007), by the end of Year 4: "*Students will read, respond to, and think critically about texts in order to meet the reading demands of the New Zealand Curriculum at level 2. Students will locate and evaluate information and ideas within texts appropriate to this level as they generate and answer questions to meet specific learning purposes across the curriculum."*

Moreover, the National Standards sets out guidance in terms of key characteristics of texts that students should read at level 2 of the New Zealand Curriculum. For example, one of the characteristics (Ministry of Education, 2007) to meet the reading demands of the curriculum at level 2 is to have *some abstract ideas that are clearly supported by concrete examples in the text or easily linked to the students' prior knowledge*.

In the following subsection, I outline some frequently used assessment tools that are well aligned with the New Zealand Curriculum in reading, writing, and mathematics in primary schools in New Zealand.

1.1.1.2 Common assessment tools in primary education

In general, assessment tools in primary education in New Zealand cover four learning areas: reading, writing, mathematics, and Māori-medium (Ministry of Education, 2007; Ministry of Education, n.d.). While it is an important learning area, the Māori medium is out of the scope of this thesis and thus will not be elaborated further. The New Zealand Ministry of Education maps out a comprehensive summary of assessment tools available in those learning areas in New Zealand by year levels (Ministry of Education, n.d.). Table 1.1 summarises some of the most frequently used tools in reading, writing, and mathematics by year levels.

Table 1.1 Some Frequently Used Assessment Tools in Reading, Writing, and Mathematics by Year Levels in New Zealand Primary Schools

<table>
<tr><th>Year Level</th><th>Running Records</th><th>PM Benchmarks</th><th>e-asTTle</th><th>PAT</th><th>STAR</th></tr>
<tr><td>1</td><td rowspan="3">Reading</td><td rowspan="8">Reading</td><td rowspan="3">Writing</td><td></td><td></td></tr>
<tr><td>2</td><td></td></tr>
<tr><td>3</td><td>Mathematics</td><td rowspan="8">Reading</td></tr>
<tr><td>4</td><td rowspan="7"></td><td rowspan="7">Reading, Writing, Mathematics</td><td rowspan="7">Reading, Mathematics</td></tr>
<tr><td>5</td></tr>
<tr><td>6</td></tr>
<tr><td>7</td></tr>
<tr><td>8</td></tr>
<tr><td>9</td><td></td></tr>
<tr><td>10</td><td></td><td></td></tr>
</table>

1.1.1.2.1 Running Records

Running Records are used to assess students' reading progression from year 1 to 3 through systematically observing and evaluating student reading aloud from any text and in any setting (Clay, 2000; Ministry of Education, 2000). It is the most commonly used assessment tool for students in the first three years at New Zealand primary schools (Blaiklock, 2004). Teachers select books of appropriate difficulty level for each student

to read (providing adequate teaching instructions) and analyse the running records. The results of running records are used to determine students' reading achievement and progression, where low progress students after a year at school will be selected for Reading Recovery (Clay, 2000; Ministry of Education, 2000).

The Running Records uses a colour wheel scheme which indicates a student's reading age in relation to expected progression through the core ('Ready to Read') reading series (Ministry of Education, 2009). The colour wheel grading of 24 levels begins at 'magenta' which is aligned to an early level 1 in reading of the New Zealand Curriculum and ends at 'silver' after the end of year 3, which is aligned to level 2 in reading of the New Zealand Curriculum.

1.1.1.2.2 PM Benchmarks

The PM Benchmarks (PM Benchmarks, n.d.) was developed as a reading assessment resource for students from year 1 to 8. It includes two kits, each consisting of 46 levelled texts ranging progressively from emergent level to reading age of 12. Each text is administered to students as unfamiliar text, and the benchmarks are designed to evaluate silent reading for oral retelling, oral reading for miscue analysis and comprehension to ascertain the student's level of understanding (PM Benchmarks, n.d.). The PM Benchmark aims to measure student reading accuracy, fluency, behaviour, and comprehension through a calculation of reading age using reading level tables, and their ability to answer comprehension questions (PM Benchmarks, n.d.).

The PM Benchmarks uses a colour wheel scheme of 30 levels for students up to age 12. The PM Benchmark levels 1 to 24 are equivalent to those of Running Records for students from year 1 to 3 (Ministry of Education, 2009). The Ministry of Education website (Ministry of Education, n.d.) offers colour wheel schemes under each tool and their equivalent New Zealand Curriculum levels, as well as the expected reading age and its expected progressions.

1.1.1.2.3 e-asTTle

e-asTTle is an online standardised assessment tool, developed to assess students' achievement and progress in reading, mathematics, writing, and in pānui, pāngarau, and tuhituhi. The reading and mathematics assessments have been developed primarily for students from year 5 to 10, but because they test curriculum levels 2 to 6 they can be used for students in lower and higher year levels (NZCER, 2012). The e-asTTle writing tool has been developed for the assessment of students from years 1 to 10 (NZCER, 2012).

The e-asTTle tool can be administered to either individuals or groups, although more commonly used with groups of students (NZCER, 2012).

The e-asTTle reading assessment aims to assess student reading comprehension through a combination of multi-choice questions and open questions where the proportion of open questions is determined by the teacher (NZCER, 2012). It evaluates reading capability in a range of deep and surface features including audience and purpose, content and ideas, structure, and language resources, as well as grammar, punctuation and spelling (NZCER, 2012). The reading tool provides e-asTTle reading scores overall and for each level of the features.

The e-asTTle writing assessment is designed to assess student writing capability within five different writing categories including the recount, describe, explain, narrate, and persuade (NZCER, 2012). Each writing test is scored by teachers against rubrics of seven elements of writing including ideas, structure and language, organisation, vocabulary, sentence structure, and punctuation (NZCER, 2012). Rubric scores set by the marker are then converted to e-asTTle writing scores overall and for each level of the elements.

The e-asTTle mathematics assessment aims to assess student mathematical ability covering categories including number knowledge and operations, algebra, geometric knowledge and operations, measurement, and probability and statistics (NZCER, 2012). Each mathematics test can cover up to three of the above categories and consists of a combination of multi-choice questions and open questions where both the choice of categories and the proportion of open questions are determined by the teacher (NZCER, 2012). The mathematics tool provides e-asTTle mathematics scores overall and for each category.

e-asTTle also provides comprehensive interpretations and specific feedback that relate to student achievement and progress in reading, writing, and mathematics and further identifies areas of student weakness and strength (NZCER, 2012). It identifies curriculum outcomes as e-asTTle scores (in reading, writing and mathematics) are well aligned with curriculum levels. Student progress can be tracked over time, and their achievement and progress can be compared with national norms (NZCER, 2012).

1.1.1.2.4 PAT (Progressive Achievement Test)

The PAT (Progressive Achievement Test) tools are standardised assessment tests designed by the New Zealand Council for Educational Research (NZCER) team that help teachers to assess students' ability in a variety of learning areas including reading and

mathematics (Darr, Neill, Stephanou, & Ferral, 2007; Darr, McDowall, Ferral, Twist, & Watson, 2008).

The PAT Reading Comprehension tool (Darr, et al., 2008) is designed to determine student achievement level and progress in reading comprehension, and to indicate possible gaps in student learning. It includes seven tests designed for students from Years 4 to 10. Each test consists of only multi-choice questions on seven to eight extended texts, which include narrative texts, poems, reports, explanations, and procedural and persuasive texts (Darr, et al., 2008). Each test question can be categorised as retrieval questions, local inference questions, or global inference questions (Darr, et al., 2008).

The PAT Mathematics tool (Darr, et al., 2007) aims to determine student achievement level in the knowledge, skills, and understandings of mathematics. It includes nine tests designed to be used at year 3 to 10. Each test is organised into four categories including number knowledge, number strategies, geometry and measurement, and statistics, while Tests 5–7 include an additional category named algebra (Darr, et al., 2007).

PAT provides raw scores measuring the absolute number of correctly answered questions, which are converted to scale scores to indicate the location of achievement on a described equal-interval scale (Darr, et al., 2007; Darr, et al., 2008). As a result, PAT scale scores can be compared directly between year levels and across assessment tests within the assessment tools. Scale scores can also be used to compare a student's achievement with the achievement of the normative samples at different year levels (Darr, et al., 2007; Darr, et al., 2008). Moreover, scale scores can be used to estimate the curriculum level a student is working at by comparing the student's scale score with the expected curriculum levels set by the New Zealand Curriculum (Ministry of Education, n.d.). Another transformation of the PAT scale scores is the stanine scores where scale scores are scaled on a nine-point standard scale with a mean of five and a standard deviation of two (Darr, et al., 2007; Darr, et al., 2008).

1.1.1.2.5 STAR (Supplementary Test of Achievement in Reading)

The STAR (Supplementary Test of Achievement in Reading) tests are another nationally normed standardised assessment of reading designed by the NZCER. It measures decoding of familiar words, sentence comprehension, paragraph comprehension and vocabulary for students from year 3 to 9 (Elley, Ferral, & Watson, 2014). From Year 7 onwards, STAR has additional measures on the language of advertising and reading different genres or styles of writing. The STAR has parallel forms and can be given at

three points during the school year. Similar to the PAT tool, the STAR results are reported as both raw scores, stanines and scale scores (Elley, et al., 2014). STAR scale scores are the most appropriate for monitoring and comparing student achievement over time and between groups as scale scores are independent of year level, test, and time of year when the test was administered. The STAR also allows teachers to compare achievement to national norms and diagnose specific reading needs such as needs in decoding or paragraph comprehension. (Please note that the STAR has subsequently been revised – what we have reported here is the version used at the time of this study).

Table 1.2 Key Measures of Frequently Used Assessment Tools in Reading, Writing, and Mathematics in New Zealand Primary Schools

Feature	Running Records	PM Benchmarks	e-asTTle	PAT	STAR
Key Measures	Reading levels	Reading levels	aRs, aWs, aMs	patr, patm	star
Type	Non-standardised	Non-standardised	Standardised	Standardised	Standardised
Validity	Some evidence	Some evidence	Tested	Tested	Tested
Reliability	Not available	Not available	Tested	Tested	Tested
Scaling	Unequal intervals	Unequal intervals	Equal intervals	Equal intervals	Equal Intervals
Measurement Error	Not available	Not available	± 40 units	± 4 units	± 4 units
Norms	Not available	Not available	Quarterly, by year level	At year start, by year level	At year start and end, by year level
NZC	Level 1 to 2	Level 1 to 4	Level 1 to 6	Level 1 to 5	Level 1 to 5
Interpretations	Criterion-referenced	Criterion-referenced	Norm-referenced	Norm-referenced	Norm-referenced

1.1.1.3 Common assessment measures in primary education

Table 1.2 summarises the key measures from these common assessment tools discussed in the previous subsection.

Each key measure as a potential candidate for the valued student outcome of a large-scale intervention study is discussed in detail in the following subsections by the following features: (a) test type (i.e., standardised or non-standardised test); (b) testing validity and reliability; (c) scaling (of equal or unequal intervals); (d) measurement errors; (e) its norms of a nationally representative sample; (f) alignment with the New Zealand Curriculum; and (g) score interpretations (i.e., criterion-referenced or norm-referenced).

1.1.1.3.1 Validity and reliability

Assessment validity and reliability are two critical features to consider whether a test and subsequently its measure should be considered as the valued student outcome for models in longitudinal intervention studies (Briggs, 2012). Assessment validity describes the extent to which the assessment tool measures what it was designed to measure. Assessment reliability describes the extent to which the assessment measures learning consistency between repeated measures.

As indicated in Table 1.2, Running Records and the PM Benchmarks are non-standardised assessment tools. While there is some evidence in testing validity for the Running Records and the PM Benchmarks, as texts have been reviewed and levelled through extensive trialling, there is lack of evidence in testing and re-testing reliability (Blaiklock, 2004).

e-asTTle, PAT, and STAR are standardised tests that are administered and scored in a consistent, predetermined, and standard manner considering factors including the questions, conditions for administering, scoring procedures, and interpretations (Darr, et al., 2007; Darr, at al., 2008; Elley, et al., 2014; NZCER, 2012). All items under each e-asTTle, PAT, and STAR test were thoroughly examined for validity and reliability. For example, the STAR test-retest reliability r was 0.91 and the split-half reliability $r = 0.91$ for the total scores. The e-asTTle writing assessment has a student reliability index of 0.96, a prompt and marker reliabilities of 0.99, and an element reliability of 1.00 (NZCER, 2012). Thus, these tests are a valid and reliable indicator of student achievement in each of the tested areas.

1.1.1.3.2 Scaling

A thorough understanding of the scaling of a valued student outcome is essential to the development of value-added models for intervention evaluation (Briggs, 2012). It serves as a data cleaning protocol in detecting missing values and outliers, as well as a foundation for data modelling in determining the appropriate statistical methods and checking for statistical assumptions.

The e-asTTle scales (in reading, writing and mathematics) are based on the Multifacet Rasch Model (MFRM), which extends the Rasch Measurement Model (RMM) by considering facets including student proficiency, item difficulty, marker severity, and the difficulty of the prompt (NZCER, 2012). It is consistent between tests and revisions, which ranges from approximately 700 to 2000 units with a standard deviation of 100 units based on randomised national samples (NZCER, 2012).

Tests developed by NZCER (PATs and STAR) uses scale scores based on Item Response Theory (IRT) by considering only the difficulty of the questions to report student achievement and progress (Darr, et al., 2007; Darr, et al., 2008; Elley, et al., 2014). These scales can be used to describe the relative difficulty of every question in each test. As a result, it can be described in terms of the knowledge and skill associated with questions at different scale locations (Darr, et al., 2007; Darr, et al., 2008; Elley, et al., 2014). Each PAT and STAR scale ranges from approximately 5 to 105 units with standard deviations of 6 for RAT mathematics, 8 for PAT reading, and 8 for STAR. Distributions of each scale score by year groups and test type are provided by the NZCER (Darr, et al., 2007; Darr, et al., 2008; Elley, et al., 2014).

Both e-asTTle and the NZCER designed scales have equal intervals (Darr, et al., 2007; Darr, at al., 2008; Elley, et al., 2014; NZCER, 2012). It means that they represent the same amount of progress no matter where the student is located on the scale. Take PAT mathematics, for example, a student who improved from 5 to 15 units is considered to have made an equal amount of learning *progress* as a student who moves from 85 to 95 units on the same scale. On the other hand, intervals between reading levels (of both the Running Records and the PM Benchmarks) are not equal in terms of progress (Ministry of Education, 2009). For example, a year 2 student is expected to improve by 3 levels from level 14 (‘Green’) to level 17 (‘Turquoise’) in one academic year, while an equal interval of 3 reading levels is expected to take two academic years for a year 6 student.

The scale units and standard deviations, in general, are all relative to each scale and can be further normalised into a common scale such as stanine scores and z-scores. The stanine score follows an asymptotically normal distribution with a mean of 5 and a standard deviation of 2, while the z-score is the standard normal variable with a mean of 0 and a standard deviation of 1.

1.1.1.3.3 Measurement error

Measurement error is a key feature to consider when comparing scale scores in terms of valued student outcomes (Briggs, 2012). Each scale score discussed in the previous subsection is provided with measurement errors that incorporate components such as inconsistency of student response (Darr, et al., 2007; Darr, at al., 2008; Elley, et al., 2014; NZCER, 2012). One key characteristic of IRT-based scale scores is that its measurement errors follow a U shape in which the error is greatest for students with either very high or very low scores (Darr, et al., 2007; Darr, at al., 2008; Elley, et al., 2014; NZCER, 2012). As a result, it is unwise to ignore measurement errors when applying scaled scores as the valued student outcome for achievement levels. Take e-asTTle writing, for example, it reports the measurement error as a plus or minus (±) range, e.g., 1450 ± 40 units (NZCER, 2012). Consider two students who scored 1450 ± 40 units and 1360 ± 60 units respectively. While the initial scores are different, it would be incorrect to classify the 'true' writing achievement levels as different, because between them as the score ranges overlap.

1.1.1.3.4 Norms and New Zealand Curriculum

Assessment norms of quality and clarity based on well-designed and randomised national representative samples are vital to educational intervention studies, of which most follow quasi-experimental design (Lai, et al., 2014; Lai, et al., 2018; Jessen, et al., 2018a; Jessen, et al., 2018b). A quasi-experimental design shares similarity with the traditional randomised experimental design, except for the random assignments of intervention or control (Morgan, 2000). Consequently, the control group in an interventional study based on a quasi-experimental design is treated as some criterion such as an assessment norm other than a random assignment to estimate the causal intervention effects (Lai, et al., 2014; Lai, et al., 2018; Jessen, et al., 2018a; Jessen, et al., 2018b).

Another natural criterion in New Zealand primary education system is the New Zealand Curriculum level as all assessment tools used are required to make direct alignments to the New Zealand Curriculum, as has been discussed in the previous two subsections.

e-asTTle provides norms and curriculum levels expected for each subject by year level and by a quarter of the year. These norms indicate how students on average in New Zealand are progressing in the subject based on a national sample of over 150,000 students in each of reading, writing, and mathematics (NZCER, 2012). The expected curriculum level, on the other hand, indicates how students on average in each year level should be progressing, according to the specifications and requirements set by the New Zealand Curriculum (Ministry of Education, 2007; NZCER, 2012). e-asTTle aligns an interval of scale to one of the three performance bands (i.e., basic, proficient, and advanced) at a curriculum level (NZCER, 2012). Take writing, for example, scores in the range of 1423 to 1459 units are categorised as "3-basic" (3B). It means that students scoring within this range are writing at an early level 3 of the New Zealand Curriculum.

NZCER provides distributions of scale scores by year levels based on national reference samples for both PAT (at year start, i.e., Term 1) and STAR (at both year start and year end, i.e., Term 1 and 4) (Darr, et al., 2007; Darr, et al., 2008; Elley, et al., 2014). These tests are well aligned with the New Zealand Curriculum, as each question of each test is associated with a curriculum level and a scale unit (Darr, et al., 2007; Darr, et al., 2008; Elley, et al., 2014). It means that students' scale scores can estimate their curriculum levels. Take PAT mathematics, for example, considering a student who scored a scale score of 50 units, he or she is expected to be able to answer approximately half of the questions in PAT Mathematics that are considered to represent level 3 of the New Zealand Curriculum.

Table 1.3 summarises the alignments of those key achievement measures in reading, writing, and mathematics to the New Zealand Curriculum levels (Ministry of Education, n.d.). Mean scale scores at year start (i.e., Term 1) are used for PAT reading, mathematics and STAR reading. Mean overall scores at year end (Term 4) from e-asTTle are used for reading, writing, and mathematics. The term "*Early*" used in the description of the New Zealand Curriculum levels means that a student is expected to work toward that level. National norms come from testing done at the *beginning* of the school year, so if students sit PATs towards the end of the year, it is best to compare their achievement against the national reference group for the next year level (Darr, et al., 2007; Darr, et al., 2008).

It is apparent from Table 1.3 that the alignments to the New Zealand Curriculum levels are necessary and useful. As discussed earlier, schools in New Zealand have the flexibility in choosing any assessment tools that they desire, which in practice lead to a variety of

achievement measures available within and between schools across the three learning areas.

Table 1.3 Alignments of Key Achievement Measures in Reading, Writing, and Mathematics to the New Zealand Curriculum

Year Level	Reading					Writing	Mathematics		NZC
	Running Records	**PM Benchmarks**	**PAT**	**STAR**	**aRs**	**aWs**	**aMs**	**PAT**	
1	Green | L14	Green | L14				1091			Early Level 1
2	Turquoise | L17	Turquoise | L17				1249			Level 1
3	Gold | L22	Gold | L22		53.8		1342		21.4	Early Level 2
4		Silver | L23	28.8	81.4	1333	1407	1389	30.6	Level 2
5		Emerald | L25	35.8	97.6	1390	1458	1430	38.9	Early Level 3
6		Emerald | L26	45	109.0	1426	1500	1466	45.1	Level 3
7		Ruby | L28	53.2	117.9	1453	1535	1500	49.6	Early Level 4
8		Sapphire | L30	60.4	125.2	1494	1566	1535	55.0	Level 4
9			67	133.7	1519	1593	1567	60.6	Early Level 5
10			76.5		1567	1617	1601	65.4	Level 5

e-asTTle reading (aRs), writing (aWs), and mathematics (aMs)

1.1.1.3.5 Interpretation

It is essential to distinguish assessment measures by the way they are designed for interpretation, especially in applications to isolate the direct effect of interventions on student achievement (Briggs, 2012). Assessment tools can be categorised into two distinct types in terms of the ways test scores are used and interpreted: criterion-referenced and norm-referenced (Bond, 1996; Briggs, 2012). The Running Records and the PM Benchmarks are both designed as criterion-referenced assessments where their assessment measures (i.e., reading levels) show what students can or cannot read in relation to a specific reading age (Clay, 2000; Blaiklock, 2004). Teachers using these criterion-based tools are required to judge whether the student has achieved the expected reading age (Clay, 2000). e-asTTle, PATs and STAR, on the other hand, are all norm-referenced tests where their assessment measures (i.e., scale scores) compare student achievements with a randomised normative sample at a given point in time (Darr, et al., 2007; Darr, et al., 2008; Elley, et al., 2014; NZCER, 2012).

However, the New Zealand Curriculum aligns assessment measures of different types in the primary education system (see Table 1.3) and offers a shared criterion-referenced measure (i.e., the New Zealand Curriculum levels). It means that student achievement in reading, writing, and mathematics can be interpreted with regards to the defined curriculum, regardless of the assessment tools (Ministry of Education, 2007; Ministry of Education, 2009; Ministry of Education, n.d.).

While much of the existing studies have used the norm-referenced scale scores (or equivalent) from those standardised assessments (Lai, et al., 2014; Lai, et al., 2018; Jesson, et al., 2018a; Jesson, et al., 2018b), in some situations it is more ideal to use the criterion-referenced New Zealand Curriculum levels as the valued student outcomes in primary education. This argument will be discussed in the following subsection. Note that while the New Zealand Curriculum level and its nature of a criterion-referenced measure is worthy of academic debate, it is out of the scope of this thesis. I simply follow the given manual and instructions outlined by the Ministry of Education (Ministry of Education, 2007; Ministry of Education, 2009; Ministry of Education, n.d.).

1.1.1.4 New Zealand Curriculum levels as the measurement of valued student outcomes

Student achievement measures in New Zealand primary education include both continuous and categorical measurements. Continuous outcome variables in this context include overall scores in e-asTTle tools, raw scores and scale scores in both PATs and

STAR. Categorical outcome variables include reading levels in the Running Records and the PM Benchmarks, stanine scores in PATs and STAR and the New Zealand Curriculum levels and its related binary indicator: at/above the expected curriculum levels and below the expected curriculum levels.

The preference of choosing the New Zealand Curriculum levels or its binary indicator as the valued student outcomes in measuring intervention effectiveness based on large-scale longitudinal and cross-sectional studies involving primary schools in New Zealand, is a result of a combination of practical and theoretical concerns.

The practical concern is the consistency of measurement tools used between and within schools. The issue of varying assessment tools used between schools is apparent as schools in New Zealand decide on how they evaluate learning progress independently. Though in some learning communities, certain assessment tools are shared collaboratively, the New Zealand education system does not universally share one tool for all. Beyond that, there are also significant concerns regarding the consistency of assessment within schools between learning areas and across times. For a given year, it is prevalent in schools to use tools from different providers for different learning areas. One common combination of using STAR for reading, e-asTTle for writing, and PAT for mathematics covers three different assessment resources. Across the years in longitudinal studies, it is also common for schools to change tools within a learning area. There was a significant shift from the STAR reading assessment to the PAT alternative when they were first introduced.

Those issues lead to problems in using their scale scores as either the valued student outcomes or independent measures of prior achievements. For example, consider a study to evaluate intervention effects on reading comprehensions where STAR reading tests were used prior to the intervention and PAT reading tests were used during and after the intervention. Using scale scores from each test across intervention time (pre and post) will lead to inconsistent measurements of data for both outcome and control variables assuming that prior achievement is a key covariate of interest.

The theoretical concern is the validity of using the standardised testing measurements as the valued student outcome to measure student learning outcomes (Bond, 1996). Norm-referenced measurements such as scale scores intend to position students' ability rather than evaluate their learning outcomes as opposed to criterion-referenced measurements such as the New Zealand Curriculum levels. For example, consider two Year 4 students

both at the expected curriculums in reading (i.e. level 2 of the NZC). One student is the top student in the school well above the expectation, and the other is right at the expected level, variations between their achievements based on a norm-referenced score can be as high as two standard deviations (e.g. a difference of 4 between stanine 5 and 9 with the known standard deviation of 2). However, one can argue that their learning outcomes do not differ since they are both at or above the expected curriculum level. They have learned what they were expected to.

This significant difference in measurement variations under two different regimes is of critical concern when considering those measures to be modelled for teacher classifications, school accountability, and intervention evaluations. While it is important to retain as much information (i.e. variation) in the outcome measure to maximise statistical power for making inferences, it is equally crucial to develop statistical models within a thorough and logical theoretical framework. Arguably, the curriculum level is the primary statement of valued student outcomes as clearly stated in the New Zealand Curriculum whose constructs are essentially defined by expectations and progressions (Ministry of Education, 2007; Ministry of Education, 2009).

The choice of a criterion-referenced valued student outcome in primary education becomes more apparent when a longitudinal intervention study tracks students from primary to secondary, as the current secondary education system in New Zealand is a criterion-referenced framework called The National Certificate of Educational Achievement (NCEA). In the following subsections, I discuss the NCEA framework and its complex data structure in detail.

1.1.2 Valued student outcomes in secondary schools

1.1.2.1 The National Certificate of Educational Achievement (NCEA)

The National Certificate of Educational Achievement (NCEA) was first introduced in New Zealand in 2002 and is currently the main qualification system for secondary school students. It is on the National Qualification Framework (NQF), which is a list of qualifications approved by New Zealand Qualification Authority (NZQA). Since its introduction, it has been reviewed and improved so that it can serve well and fairly as a New Zealand secondary school qualification. Some notable changes include the introduction of endorsement of certificates and subjects in 2007 and 2008 respectively, an increased amount of external moderations on internally assessed standards (to 10% of all internal assessments) since 2008, and realignments of NCEA level 1, 2, and 3

standards in 2011, 2012, and 2013 respectively (Harris, 2007; Madjar & McKinley, 2011).

There are three different levels of NCEA certificates: Level 1, 2, and 3, which in general are attained by students in years 11, 12 and 13 respectively. Each certificate is achieved when students have met its requirements (see Table 1.4 below).

Table 1.4 List of Requirements for NCEA Certificates and Awards

NCEA Qualification and Award	Requirements
Level 1	At least 80 credits in total, at least 10 credits in Literacy and at least 10 credits in Numeracy
Level 2	At least 60 credits at Level 2 or above, at least 80 credits in total, achieve Level 1 Literacy and Numeracy
Level 3	At least 60 credits at Level 3 or above, another 20 credits at Level 2 or higher, achieve Level 1 Literacy and Numeracy
UE	Achieve NCEA Level 3, achieve at least 14 Level 3 credits in at least 3 approved subjects, 8 credits (4 in reading and 4 in writing) in English or te reo Maori at Level 2 or higher, 14 credits in Numeracy at Level 1 or higher

University Entrance (UE) is the minimum requirement to enter any tertiary institutions in New Zealand. It is awarded to students with NCEA Level 3 certificates who have met the specified UE requirements, as described in the table above. However, different courses at different universities may set additional requirements in addition to UE; for example, the University of Auckland introduced a rank score system and use it as an independent measure for admission selections. (The University of Auckland, n.d.).

NCEA offers more than 50 courses across both academic (e.g., English, mathematics and science) and practical (such as business administration, drama, and home economic) spectra. Each subject is made up of a variety of standards, each of which describes a specific knowledge or skill and is worth a certain amount of credits. There are two types of standards, unit standards that are based on competency, and achievement standards that are based on New Zealand Curriculum (Madjar & McKinley, 2011). For example, "*Evaluate statistically based reports*" is a level 3 achievement standard worth four

credits, typically assessed in the subject Mathematics and Statistics. An example of unit standards is "*Demonstrate knowledge of enterprising behaviour, innovation and entrepreneurship in business contexts*", which is a level 1 standard for Business Studies and it is worth two credits. Teachers at school assess unit standards internally during the year. Some achievement standards are assessed in the same fashion; however, most of them are assessed externally through examinations or portfolios at the end of the year by NZQA.

There are two grades available to unit standards: *Not Achieved* and *Achieved*, while achievement standards are graded with two additional endorsements: *Merit* and *Excellent*. NCEA adapts a standard-based system where students will attain a standard and earn credits towards that standard if they have satisfied the standard's requirement regardless of the other students' performance. As a result, this creates a set of criterion-referenced achievement data in the form of categorical (ordinal) data.

1.1.2.2 NCEA achievement data and their aggregations and statistics

NCEA achievement data has a complex structure which can be described by two interactive dimensions: achievement levels and achievement measures. Achievement levels include standard, course, and qualification in the ascending order of level of aggregations. It means that under the NCEA system, standards are nested within courses, and courses are nested within qualifications. Achievement measures can be categorised into credits, attainment, and endorsement, which are available at each achievement level. Credits are a continuous form indicating the number of credits a student achieved in a standard, a course and a qualification. Attainment is a binary outcome variable indicating whether a student attained a standard, a course certificate, or a qualification certificate. The endorsement is an ordinal measure which extends attainment into three categories of performance: excellent, merit, and achieved (NZQA, 2009).

NCEA achievement data are available in the form of either raw data as collected by each secondary school in New Zealand, or aggregated data and statistics.

1.1.2.2.1 Raw data

NCEA raw achievement data includes, for each student, a list of all attained NCEA qualifications, courses, and standards, as well as any endorsements awarded, and a number of credits cumulated for each attained qualification, course, and standard. An example of a data report is Record of Achievement that each student receives at the

beginning of each year where he or she had credits reported to NZQA in the previous year.

NCEA raw achievement data are rich, comprehensive, as well as private and confidential. However, such data can be obtained with consents from individual students and schools (provided these data will be used ethically and responsibly).

1.1.2.2.2 Aggregated data

There are many types of aggregated data and statistics available in the NCEA system, and all these transformations have different implications for interpretation and further analysis. While all these forms of data have useful functions within themselves, it is essential to understand the differences and the implications in interpreting student outcomes based on these aggregated data.

We can aggregate achievement data within and between qualification levels, standards, and courses, as well as a combination of them. For example, the number of credits obtained in a selection of level 2 standards with high literacy requirements from English, Biology, and Science (collectively called *SLAS*, see Table 4.2) can be aggregated to measure Year 12 students' ability in critical literacy. Alternatively, attainment rates of mathematics standards (i.e., the number of standards attained divided by the number of standards participated) for each level can be used to track students' progress in mathematics from year 11 to 13.

We can also aggregate achievement data through data transformations. The University of Auckland employs a rank score system to grade students, relative to one another. This process converts rich and comprehensive student achievement data within the approved subject standards to a single continuous value as a mean of admission requirements (The University of Auckland, n.d.). This scoring system allocates 0 points for "not achieved", 2 points for "achieved", 3 points for "merit", and 4 points for "excellence". Similarly, Michael Johnston (Ministry of Education, 2012) developed the NCEA level 3 achievement score for a comparable purpose, which rates students' grades in their NCEA level 3 qualifications against other students in the same year. It transforms the criterion-referenced NCEA level 3 data to a norm-referenced score between 0 and 100. This measure has been used in some studies that assess students' success in tertiary education in relation to their achievement at NCEA level 3 (Scott, 2008; Shulruf, Hattie, & Tumen, 2008). Robinson, Bendikson, McNaughton, Wilson, and Zhu (2017) introduced a

benchmark indicator based on aggregated statistics to measure leadership and organisational performance over time.

1.1.2.2.3 Norms

NZQA also provides some aggregated data and statistics, which can be used as norms for intervention studies based on quasi-experimental designs. These aggregated statistics are available both nationally and at the level of individual schools. At the national level, aggregated datasets and statistics are further broken down by decile band, gender, and ethnicity, while aggregations at school level are broken down by only gender and ethnicity (as decile is implicit within the school).

NZQA's reporting of aggregated data and statistics has changed over the years. From 2004 until 2013, aggregations were presented in the forms of both numbers and percentages. However, there were concerns that individual students may be identifiable from the numbers – particularly in schools with small roll where there may only be a few students participating in any given level. Thus, since 2014, aggregated data have only been presented as percentages (NZQA, 2014).

National attainment rates of a standard or qualification are calculated based on two different cohorts, which are used as the denominators: school roll cohort and participating cohort (NZQA, 2009). School roll cohort uses the student population (given by each school in July of the academic year being reported), while the participating cohort includes only students who were considered to be attempting to achieve an NCEA qualification in the year. In general, school-roll based percentages underestimate achievement rates from a school or nationally as not all students may be aiming to obtain the associated NCEA qualification in that particular year. However, students do not register or enrol in NCEA qualification, and as a result, there are no exact figures of the number of participating students. So, in each year the 'participating' cohorts are estimated by the number of students who have attempted to achieve the minimum amount of credits during a year for each NCEA qualification. For example, a student will be considered as a participant of NCEA Level 1 qualification if he or she has entered standards of Level 1 or above that are worthy of 80 credits or more in total.

Moreover, there are two types of achievement rates calculated at both national level and school level: (a) current year achievement rate which was calculated based on those students who have attained the qualification of interest within the academic year being reported; and (b) cumulative achievement rate which includes all students who have

attained the qualification of interest by the end of the year being reported. For example, if a school has 100 students in Year 11, of which 70 attain NCEA Level 1 in 2010, then the current year achievement rate of attaining NCEA Level 1 qualification in 2010 by Year 11 students is 70%. Let us assume all those 100 students stayed in the same school in the following year, and 10 of them attained NCEA Level 1 in 2011. In this case, the current year achievement rate of attaining NCEA Level 1 qualification in 2011 by Year 12 students is 10% in 2011, and the cumulative achievement rate of Level 1 attainment is 80%. Note that, though in general cumulative achievement rates are higher than current achievement rates in the percentage of attainments, it is not necessarily true that a higher percentage of endorsements are awarded (NZQA, 2009) – that is, more time to complete a qualification does not necessarily confer an advantage in terms of endorsements.

It is suggested to use school roll-based current year data as norms, as it is more consistent and reliable particularly for longitudinal studies (Hodis, Meyer, McClure, Weir, & Walkey, 2011; Meyer, et al., 2009; Meyer, Weir, McClure, Walkey, & McKenzie, 2009). As indicated previously, the intention of students to participate in NCEA qualifications cannot be directly measured without formal registration or enrolment. Estimation methods currently in use to identify 'participating' students rely on the students attempting the minimum threshold for the number of credits at different NCEA levels (e.g., 80 credits at NCEA level 1). Furthermore, current year achievement rates are better data source to match the underlying expectations that students should attain NCEA level 1, 2, and 3 at Year 11, 12, and 13 respectively.

1.1.2.3 Challenges of NCEA achievement data

While the flexibility and complexity of the NCEA system cater to the needs of an increasingly diverse student population and supports greater success in a variety of course combinations at varying levels, it is these very characteristics that lead to both practical and theoretical challenges when using NCEA achievement data for evaluating valued student outcomes.

The practical challenge is the issue of inconsistency in NCEA achievement measures between schools and academic years. The NCEA system allows each school to design their curriculum, subject to a combination of factors including school goals, student needs, teaching resources, and school facilities. As a result, standards, courses, and qualifications offered vary between schools, as well as between academic years within each school. For example, consider the case of the aggregated number of credits obtained

in *SLAS* that is worth 24 credits; the added availability of a standard worth six credits would cause significant variations.

The theoretical challenge is the construct of a valid and reliable measure based on NCEA achievement data. It is important to consider what are 'valued student outcomes', and what constitutes the achievement of such, within the context of the research to be undertaken. For instance, consider again the case of the aggregated number of credits obtained in *SLAS*. While such aggregation might serve well as a valuable outcome for a year 12 literacy intervention study, it is not an adequate outcome measure for a school-wide leadership intervention study.

Researchers must consider which achievement levels (standard, course, or qualification), assessment measures (credit, attainment, or endorsement), and transformations are most appropriate as the valued student outcomes to answer their research questions. For example, transformations such as the University of Auckland rank scores are well suited for research apt to tertiary education. Because they are used to determine entry, likely success, and for making judgments that have some predictive validity, they confine their attention to student achievement in traditional academic subjects or standards. However, this contrasts with one of the motivations for introducing NCEA to New Zealand secondary schools - which was to acknowledge and encourage diversities in student abilities and interests (Harris, 2007; Madjar & McKinley, 2011).

1.1.2.4 Recommended forms of NCEA data as a valued student outcome

In this thesis, NCEA raw achievement data of attainment and endorsement in standards, courses, and qualifications are the preferred valued student outcome for longitudinal intervention studies in New Zealand secondary schools. The omission of achievement measure in the number of credits and the choice of raw data are a result of minimising inconsistency raised from the challenges discussed in the previous subsections.

Credits in NCEA achievement data are designed as measurements of standard difficulty rather than student ability. As discussed as a practical issue in the previous subsection, variations in credit-based measures can be significantly influenced by the standard availability. As a result, only attainment and endorsement are considered as a valued-outcome measure for student achievement.

While the various aggregations and transformations do serve their purposes, little research has been done to prove their validity and reliability in assessing student learning outcomes. As discussed as a theoretical concern in the previous subsection, aggregations

and transformations are subject to researchers' judgements and considerations, without a common protocol. As a result, raw data is preferred to minimise inconsistency between researcher-made aggregated or transformed measures.

In summary, the valued student outcomes for intervention evaluation in the context of New Zealand education system are established in this section and they are all categorical data, which include the New Zealand Curriculum levels and its binary equivalents for primary education and the NCEA raw achievement data of attainment and endorsement in standards, courses, and qualifications for secondary education.

1.2 Value-added models

Given a stable and relevant valued student outcome, intervention evaluation can be improved through a robust and thorough model specification. In this section, I provide a general review of value-added models in the context of educational research. First, the value-added model as a means to measure intervention effects is introduced in subsection 1.2.1. Subsection 1.2.2 reviews its application in educational research, followed by a discussion on its limitations in isolating intervention effects in subsection 1.2.3. Subsection 1.2.4 provides a comprehensive review of the application of multilevel models in the value-added models for educational evaluation, followed by a discussion on robust model specifications for the multilevel models in subsection 1.2.5.

1.2.1 Value-added models as a means to measure effects

Scherrer (2011) reviewed and summarised four measurement models commonly used in assessing school effectiveness: status models, cohort-to-cohort models, growth models, and value-added models (VAMs). A status model describes student achievement at one particular time point and is useful in answering questions such as *"What is the NCEA level 1 attainment rate of school X in 2011?"* Cohort-to-cohort models compare two status models for a particular time and attempt to answer between-groups questions such as *"What is the difference in NCEA level 1 attainment rates between male and female students from school X in 2011?"* Growth models track student achievements over a period and study trajectory or trend in achievements, and answer questions such as *"What is the change in achievement rates for students from School X across 2011 to 2018?"* (Raudenbush & Bryk, 2002).

In contrast, value-added models attempt to control factors at different levels (student and school levels, and in some cases classroom, year level, or departmental levels) and

measure the proportion of growth in achievement that is as a result of an intervention-effect, a school-effect, or a teacher-effect (Sanders & Horn, 1998; Gordon, Kane, & Staiger, 2006; Harris, 2009). In other words, value-added models describe *accelerated and isolated* effects. As such, they are of particular interest in evaluation studies, especially in an education context where interventions are based on quasi-experimental designs with no random assignments (Finch & Cassady, 2014; Lai et al., 2014; Lai, et al., 2018; Jesson, et al., 2018a; Jesson, et al., 2018b).

Each of the models summarised above has its advantages and disadvantages. In this thesis I will confine my attention to value-added models as they are widely used to estimate the schools', teachers' or interventions' effects on student achievements controlling for such things as initial achievement levels, student demographics, and teaching experience (Goldstein, 1997; Goldstein, 2001; Goldstein, Burgess, & McConnell, 2007; Heck, 2000; Hill & Goldstein, 1998; Raudenbush, Bryk, & Ponsiciak, 2003; Strand, 2003; Thomas, Sammons, Mortimore, & Smees, 1997).

1.2.2 Value-added models in educational research

Value-added models were first introduced in the 1970s (Hanushek, 1971) and had recently gained prominence in educational research (Hanushek, 1992; Sanders & Horn, 1998; Goldhaber & Hansen, 2010). Comprehensive reviews of the value-added models are well documented (Braun, Chudowsky, & Koening, 2010; Briggs, 2012; Hanushek & Rivkin, 2010; Harris, 2009; McCaffrey, 2003; McCaffrey, Han, & Lockwood, 2009), and well-known associations also emphasise guidance of its application in educational research (AERA, 2015). Due to the scientific and technical limitations of value-added models, there are significant concerns of using value-added models to make *direct causal inferences* about the effectiveness of teachers, schools, and interventions (Baker, Barton, Haertel, & F, 2010; Braun, et al., 2010; Ishii & Rivkin, 2009; McCaffrey, 2003; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Raudenbush, 2009; Reardon & Raudenbush, 2009; Briggs, 2008; Briggs & Domingue, 2011; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Sanders, Wright, Rivers, & Leandro, 2009). However, Glazerman et al. (2010) argued that value-added models, despite their limitations and deficiencies, made improvements relative to other approaches to educational evaluation. Several studies were also supportive of the validity of teacher and intervention effects estimated using value-added models (Kane & Staiger, 2008; Koedel & Betts, 2011; Jesson, et al., 2018a; Jesson, et al., 2018b).

It is critical for educational researchers, practitioners, and policy-makers to understand issues with current value-added models and the rationale for what can be estimated and what cannot (Raudenbush, 2004; Briggs, 2008; Briggs & Domingue, 2011; Gelman, 2004; Ballou, Sanders & Wright, 2004; McCaffrey, et al., 2004; Lockwood, McCaffrey, Mariano & Setodji, 2007). While a value-added model, when combined with other information (such as students' growth curves), can be useful as an indicator or descriptive measure, it has limitations to measure the effectiveness of instructional practice (Briggs, 2008; Hill, Kapitula, & Umland, 2010).

1.2.3 Limitations of value-added models in measuring effects

One of the limitations is the stability of estimating effects (Simon, Erikan, & Rousseau, 2012). At the student level, measurement errors in the valued student outcomes and the inconsistencies in assessment measures over time will lead to loss of statistical accuracy in making direct inference on intervention effects (Fuller, 1987; Buonocarsi, 2010). It is of significant concern in New Zealand as schools have the agency to choose and change assessment tools at their convenience (Wilson, McNaughton, & Zhu, 2017). At the teacher or school levels, it is argued that variations in estimates of value-added can be partially explained by chance for the reasons that some cohorts of students are just "better" than others (Kane & Staiger, 2002; McCaffrey, Sass, Lockwood, & Mihaly, 2009).

Another limitation is the lack of statistical power, even in large-scale well-designed studies (Kane & Staiger, 2008; Rothstein, 2010). It is related to sample size. Rothstein (2010), for example, criticised the findings of Kane & Staiger (2008) and argued that their model did not have enough statistical power to measure discrepancies between the obverted effects and the estimated (by the value-added models) because the experiment was carried out in a relatively small sample of classrooms and teachers. Raudenbush et al. (2003) showed that precision in estimating teacher effects is modest even with large sample sizes (more than 50,000).

The omission-variable bias as a result of omitting confounding variables that lead to biased and inconsistent estimations is another critical issue of value-added models. Rothstein (2010) introduced an inspiring "*falsification*" check when intervention in the future is found by a value-added model to affect student achievement gains in the past (i.e., before the implementation of such intervention). One plausible explanation for this illogical estimation is that such estimation is the collective impact of one or more omitted

confounding variables that are correlated with the intervention. The value-added model captures the effect of those omitted confounding variables but incorrectly attributed to a future intervention. As a consequence, Rothstein (2010) argued that such omission-variable bias would also persist in the estimated effect of the current intervention. Hodis et al. (2011) addressed the latent variable of student motivations in a growth mixture model to detect different growth trajectories in achievements. In most interventions, and especially in a New Zealand context, there are possibly multiple unobserved subgroups in the student population who could respond to the intervention with different growth trajectories.

Other limitations in value-added models include determining which background factors to account for (Heck, 2000). There is no consensus in the literature on what specific components contribute to student achievement, and much depends on the makeup of the student population under scrutiny (Raudenbush & Bryk, 2002).

However, value-added models are essential in large-scale repeated assessments in education for evaluation purposes (Chudowsky, Braun, & Koenig, 2010; McCaffrey, et al., 2003; Strathdee & Boustead, 2005; Kelly & Downey 2010; Leckie & Goldstein, 2009). While the question of whether value-added models can be used to measure direct impact and make causal inferences about intervention effectiveness remains open, one can argue that robust statistical modelling such as multilevel models can significantly enhance the validity, reliability, and efficiency of value-added models (Briggs, 2008; Briggs & Domingue, 2011).

1.2.4 Application of multilevel models in value-added models

In general, value-added models apply multilevel models controlling for covariates observed at different levels (such as student-level, teacher-level and school-level as described above) to estimate the overall effects of interest (Raudenbush & Bryk, 2002; Gelman & Hills, 2007; Goldstein, 1997; Raudenbush, 2004; Timmermans, Doolaard, & De Wolf, 2011; Foley & Goldstein, 2012; Ray, 2006; Sammons, Nuttall, & Cuttance, 1993). Such applications are especially prevalent in fields such as educational and social science study (Baker et al., 2010; Braun, et al., 2010; Ishii & Rivkin, 2009; McCaffrey et al., 2003; Newton, et al., 2010; Raudenbush, 2009; Reardon & Raudenbush, 2009; Briggs, 2008; Briggs & Domingue, 2011; Braun, 2005; McCaffrey, et al., 2004; Sanders, et al., 2009), due to the natural occurrence of multilevel data structures in those research fields. Multilevel data structures arise in clustered data (which in most cases have hierarchical

structures) and also in repeated measures in longitudinal studies (Raudenbush & Bryk, 2002; Gelman & Hills, 2007; Goldstein, 1997; Raudenbush, 2004; Timmermans et al., 2011; Foley & Goldstein, 2012; Ray, 2006; Sammons et al., 1993). Gelman and Hill (2007) provide an excellent introduction to multilevel models and their applications in educational and social science study.

Co-dependency exists in observations with multilevel data structures. For example, a shared community culture can influence students within a school. Thus, it violates a critical assumption of independent observations required by most statistical models (Raudenbush & Bryk, 2002; Gelman & Hills, 2007; Goldstein, 1997; Raudenbush, 2004; Timmermans et al., 2011; Foley & Goldstein, 2012; Ray, 2006; Sammons et al., 1993). Estimates of the effects are likely to be biased when underlying co-dependency is ignored (Simon, Ercikan, & Rousseau, 2012). It, in turn, will lead to significantly misleading inferences as a result of the "*ecological fallacy*", for the reason that the distinguished within-school effects are summarised by between-school effects (Gibson & Asthana, 1998; Goldstein & Spiegelhalter, 1996; Goldstein, 1997, 2001).

Multilevel models take advantage of the existence of this co-dependency by allowing for the residual components at each level to address the violation of the underlying assumption of independent observations (in this case, students). Multilevel models measure individual-specific effects by correlations calculated based on the residual variance-covariance matrix (Raudenbush & Bryk, 2002; Gelman & Hill, 2007). For example, a multilevel model with two levels which allows for grouping of student achievement within schools would include residuals at the student- and the school-level. The variance of the student-level residuals measures the within-school effect, while the variance of the school-level residuals measures a between-school effect. It is often referred to as a random effect.

The random effects of school represent unobserved school characteristics (e.g., student motivation and parental involvement) that affect student achievement. In turn, the variance-covariance structure of the two residual variances will be estimated to measure the correlations between student achievements for students from the same school.

1.2.4.1 Multilevel model (MLM)

Before I discuss the model specifications, I would like to first justify my choice of its name in this thesis, and for the field of educational research. Statistically speaking, the multilevel model is a generalised linear mixed model (GLMM), which is an extension to

the generalised linear model (GLM) to allow for correlations between observations, as occurs, for example, in a longitudinal study (with repeated outcome variables consisting of multiple observations per individual collected over several measurement occasions (Raudenbush & Bryk, 2002)) and clustered data (where individuals are naturally nested in a higher group such as schools or regions (Raudenbush & Bryk, 2002)).

GLMMs are "*generalised*" in that they allow for outcome variables that have *generalised* distributions including normal distribution for linear regression, binomial distribution for logistic regression and Poisson distribution for Poisson regression (Raudenbush & Bryk, 2002). In some fields of research particularly educational studies, logistic regression models are frequently referred to as latent variable models with error variables following a logistic distribution (Raudenbush & Bryk, 2002). GLMMs are also "*mixed*" in that the linear predictor contains both the *fixed* (as in GLM) and *random effects*. I will discuss *fixed* and *random* effects in detail in subsection 1.2.4.2. In this subsection, it is sufficient to know that differences between groups can be modelled as a random effect.

GLMMs with application to repeated measures are often phrased as *growth-curve models* or in cases of categorical outcome variables *latent growth-curve models* (Raudenbush & Bryk, 2002). GLMMs are referred to as *hierarchical (linear) models* when they are used to model data with defined nested structures such as students nested within classrooms within schools (Raudenbush & Bryk, 2002). However, none of those names fit well when data structures become complicated, for example, in longitudinal studies of reading achievement from repeated samples over many different schools. The underlying data structure has a natural clustering in that students are nested in schools for each school term. However, student's repeated reading achievement is not only nested in each student; it is also nested within school terms. Thus, there is no streamed hierarchy but two seemingly parallel ones – in other words, it is multilevel but not hierarchical.

GLMMs are known in general variously as *random effect models*, *random coefficients models*, *mixed effect models*, and *mixed models* to place emphasis on the inclusion of *random effects* in addition to *fixed effects* (Gelman & Hill, 2007). However, those terms are statistical rather than practical. The emphasis on random effects may confuse readers and researchers in more applied fields such as education, social sciences, and medical studies. In most applied research GLMMs are used to estimate impacts, effects, differences, and added-values, all of which are fixed effects. There are also inconsistencies in the definitions of fixed and random effects (see 2.3), which can also lead to considerable confusion (Gelman & Hill, 2007).

In this thesis, I choose to use the term *multilevel model* (MLM) to refer to the collection of GLMMs that are used to model data from a sample of correlated observations due to natural clustering and or repeated measurements. As described above, this is because of the implied generalisability of the term, concerning educational research. It encompasses models with both fixed, and fixed and random effects, and shifts the focus from solely 'hierarchically-levelled' to 'multilevelled' data. The proceeding subsection will describe multilevel models in the perspective of random effects, as it is a critical concept in understanding the multilevel model specifications.

1.2.4.2 Fixed and random effects

Multilevel models are often referred to as 'mixed' models because they include both fixed and random effects (Gill, 2003). Various definitions of fixed and random effects and their differences are well documented (Gelman & Hill, 2007; Gelman, 2004; Gelman, 2006; Fox, Negrete-Yankelevich, & Sosa, 2014). Gelman (2004) argued that the terms fixed effect and random effect have variable meanings depending on the fields of study as well as the perspectives of researchers. In the following subsections, I discuss differences between the two effects from an integrated (both statistical and practical) perspective.

1.2.4.2.1 From the perspective of statistical inference

"If an effect is assumed to be a realized value of a random variable, it is called a random effect." (LaMotte, 1983)

Mathematicians and statisticians frequently use the above definition from LaMotte. In classical statistical inference (Frequentist inference) random effects are defined as categorical variables whose levels are chosen at random from a larger population (Fox, et al., 2014), for example, schools chosen at random from a list of all schools. Such a definition can be practically problematic. For example, it implies that you cannot use 'school' as a random effect when you have observed all of the schools from the population of interest. Similarly, in longitudinal studies, temporal variables can be modelled as random effects but *in practice*, repeated observations are not collected at randomly sampled time points. However, it is philosophically coherent in the sense that random effects render *inference* where fixed effects test *differences*.

When a covariate is modelled as a random effect, statistical inferences can be made about the distribution of values (i.e. the variance among the values of the outcome variables at different levels). On the other hand, when the same covariate is modelled as a fixed effect, estimated coefficients only measures the differences of values between particular levels.

Take the covariate "school" for example in an educational study. If we fit "school" as a fixed effect, we simply ask the model to estimate school-specific intercepts, α_j, for each school j. In contrast, if we fit "school" as a random effect, the model will estimate an overall intercept adjusted by "school" and a collection of school-specific random effects (adjusted for any additional school-level covariates) on the overall intercept. We can statistically test differences between fixed effects, but our interpretations will be limited to the sampled schools only. It means we are not able to make inference on schools outside of the sampled schools. Random effects, however, are normally distributed with mean 0 and an estimated variance-covariance structure. They allow us to make inference on any unobserved schools. Even in the situation of a population study where all schools are observed, random effects allow us to make the inference of any new schools that yet to be built at present.

Searle, Casella, and McCulloch (1992, Section 1.4) have a similar definition of random effects from a more practical perspective and have described the two in this way:

"*Effects are fixed if they are interested in themselves or random if there is interest in the underlying population.*"

1.2.4.2.2 From the perspective of statistical power

Random effects are estimated with *partial pooling* (or shrinkage), while fixed effects are not (Gelman & Hill, 2007). Snijders and Bosker (1999, Section 4.2) provide a more technical definition as random effects are estimated by *best linear unbiased prediction* (BLUP) in the terminology of Robinson (1991) and fixed effects are estimated using *ordinary least squares* (OLS) or, more generally, *maximum likelihood* (MLE). Estimation theory is a well-established branch of statistics and out of the scope of this thesis. Please refer to literature in statistical inference and estimation theory for further explanations on those estimators mentioned in the definition by Snijders and Bosker (1993).

From a more practical point of view, we can think of partial pooling as a way to draw information from different levels within a grouping variable (Gelman & Hill, 2007). Bates, Kliegl, Vasishth, & Baayen et al. (2015) described shrinkage or partial pooling as a mean of borrowing statistical strength. In cases where there are few observations in a group (e.g. rural schools with fewer students), the group-specific effect will be estimated partially on the more abundant observations from other groups. The fewer the observations in a group, the more information (hence statistical power) is borrowed from other groups.

Shrinkage can also be described as the way the model pulls more extreme estimates towards an overall average (Sullivan, Dukes, &Losina, 1999). The parameter estimates are brought towards the overall mean intercept and slope by shrinkage. Sullivan et al. (1999) suggested that the more extreme the estimates the greater the shrinkage.

Consider a study to evaluate school intervention success. Suppose there is a large dataset containing observations of student achievement (the valued student outcome variable) and variables arising from measures of the intervention intensity and quality (the covariates) from a sample of schools. Some schools are well represented with adequate sample sizes, but others have only a few sampled students, due to low numbers of students in any given year level. A fixed effect approach with complete pooling will ignore existing correlations in the structured data, which will, in turn, lead to biased estimation of the overall intervention effect, where schools with larger sample sizes contribute more to the overall estimation. A fixed effect approach with no pooling will yield estimations of school-specific intervention effects with varying variations, in particular, higher than expected variations due to sampling variance in poorly sampled schools. Shrinkage or partial pooling can mitigate this by pushing extreme values towards the mean across all schools. Fitting the covariate 'school' as a random effect, in this case, will render a mean overall intervention effect subjected to a statistically robust shrinkage, taking into account the unbalanced samples among schools as well as the sampling variance at the school level.

1.2.4.2.3 From the perspective of statistical assumptions

One critical difference between fixed and random effect models are the underlying assumptions about the individual-specific effects (Laird & Ware, 1982). The random effect models assume that the individual-specific effects are uncorrelated with the covariates. The fixed effect models assume that the individual-specific effects are correlated with the independent variables (Gardiner, Luo, Roman & Lee, 2009).

Consider a study to evaluate school intervention success. Suppose there is a large dataset containing observations of student achievement (the outcome variable) and a collection of covariates from a sample of schools. Figure 1 illustrates the overall correlation between the achievement and the covariate, and is positive; the higher the covariate, the higher the achievement. However, observations clustered in groups (as indicated by the different colours) presents a negative relationship, which is reasonably consistent across all groups.

In this simulated study, the "true" effect of the covariate is set at -1 and estimated effects from random effect model and fixed effect model are -0.6 and -1.15 respectively. The inconsistent and biased estimation from random effects is due to the violation of the assumption on the independence between individual-specific effects and the covariates. It is apparent in the following figure that the light-blue group clusters in the bottom left corner where the light-yellow group clusters around the top right groups. That is, there are correlations between the covariate and achievement as well as school. As suggested in the simulated result, an estimation made by a fixed effect approach is more effective and more convincing. However, this bias disappears as the number of observations within each school increases, as the weight on the fixed effect then tends to one (see e.g. Hsiao, Analysis of Panel Data, Sec. 3.3.2).

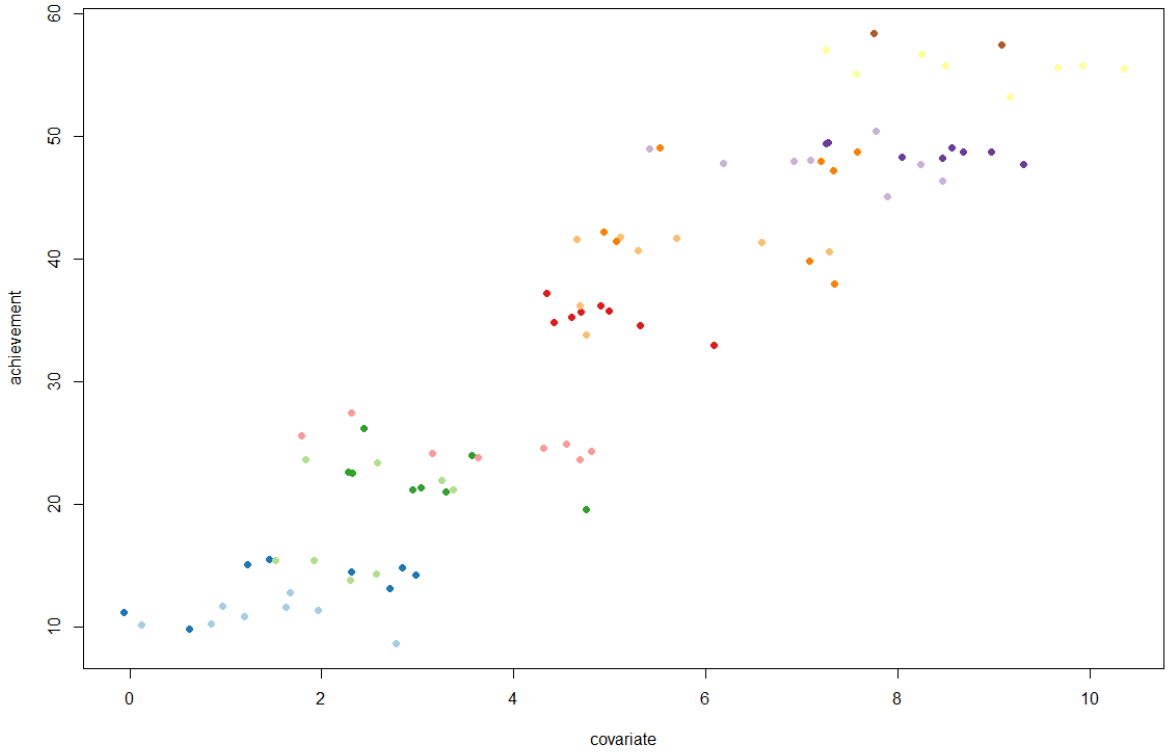

Figure 1. Influence of correlated individual-specific effects on fixed and random effects estimations.

Conceptually, it is important to assess the critical assumptions prior to model fitting. When any variable is fitted in a regression, its coefficient is estimated with the assumption of holding everything else constant, including all other observed variables in the model specification, as well as those unobserved that are not included in the model. If there are variables correlated with other variables that are not present in the model, the estimation

of its coefficient will be biased as it cannot hold that omitted variable constant. For instance, in the above example, if we omit the variable 'school' and regress the achievement over covariate in a complete pool approach, we will yield an underestimated effect which is clearly incorrect.

The fixed effects approach addresses this bias by adding the variables to the model representing the time-invariant individual-specific effects. As a result, the other coefficients in the model can be estimated, holding the individual-specific effects fixed.

In contrast, the random effects approach models the variance-covariance structure of the error terms. It does not attempt to estimate the individual-specific effects. Instead, it models the correlation between observations within groups. The estimated random effects can be seen as drifts to the overall regression line, and observations from the same groups share the same drift.

In summary, the fixed effects address the omission bias by modelling individual-specific effects while the random effect models overcome the correlation bias by modelling the covariance structure of the errors. In educational research, the correlation bias is known due to the natural clustering of students within schools; on the other hand, the omission bias is unknown as the omitted fixed effect cannot be observed.

In the next subsection, I discuss the application of a multilevel model in terms of model specifications as they are particularly critical when applied to directly evaluate intervention effectiveness, as differences in the model specifications can lead to significant differences in evaluations (Felch, Song & Smith, 2010).

1.2.5 Robust model specifications for multilevel models

Value-added models require sufficiently complex and theoretically adequate multilevel model specifications to support a valid statistical inference, particularly for the purpose of isolating effects at different levels. There is a variety of multilevel models that have been used to classify effectiveness (Goldstein, 1997; Raudenbush, 2004; Timmermans et al., 2011; Foley & Goldstein, 2012; Ray, 2006; Sammons et al., 1993). Table 1.5 outlines generalised model specifications in the ascending order of model complexity using an example of a dataset that consists of control variables at student and school-levels.

The simplest value-added model specification is the *empty model* (Goldstein, 2011) which specifies only the variance-covariance structure of the valued student outcome at student and school-level without considering any additional covariates. It is also named as the

"*type 0*" value-added models (Timmermans et al., 2011). Immediate improvement of the empty model is the "*type AA*" value-added model (Timmermans et al., 2011) which measures the learning progress made by the students controlling for their prior achievements. It is also referred to as the "*adjusted (school) comparisons*" (Goldstein, 1997). The "type AA" value-added models do not take into account contextualised information such as student demographics and school socioeconomic factors, which often lead to biased estimations of school effects (Foley & Goldstein, 2012; Goldstein et al., 2007). Consequently, contextualised value-added (CVA) models can reduce those bias by controlling for relevant factors in addition to prior achievements that may influence student learning progress (Foley & Goldstein, 2012; many authors agree on this concept, see for example Goldstein, 1997; Ray, 2006; Sammons et al., 1993). Contextualised value-added models controlling for only student-level covariates is defined as "*type A*" (Raudenbush, 2004; Timmermans et al., 2011). When both student and school-level covariates are controlled, "*type B*" is defined when school-level variables are observable (Raudenbush, 2004; Timmermans et al., 2011), and "type X" is defined when school-level variables are both observable and latent (Timmermans et al., 2011).

Model specifications are particularly critical when value-added models are applied to directly evaluate intervention effectiveness, as differences in the model specifications can lead to significant differences in evaluations (Felch, et al., 2010). In subsequent studies on the same data consisting of multiple cohorts of students taking the reading and mathematics portions of the California Standardised Test between 2003 and 2009, teacher classifications varied significantly due to differences in model specifications (Buddin, 2010; Briggs & Domingue, 2011).

Briggs (2012) credited Kane & Staiger's success (2008) in isolating the effects of teachers from other possible factors to a "*quite complex"* value-added model where its specifications controlled for both prior achievement and a collection of other variables at both the student and classroom levels (i.e., a Type B value-added model shown in Table 1.5 List of Value-Added Model Specifications Using a Two-level Structured DataTable 1.5). Briggs (2012) further argued that had the model specification been reduced to include only student-level covariates or prior achievements (such as Type AA or Type A shown in Table 1.5), by no means the success would be repeated.

Table 1.5 List of Value-Added Model Specifications Using a Two-level Structured Data

Value-Added Model	Effects	Student Level	Prior Achievement	School Level	Other Names
Type 0	Random	No	No	No	Empty model
Type AA	Mixed	No	Yes	No	Adjusted comparisons
Type A	Mixed	Yes	Yes	No	Contextualised value-added (CVA)
Type B	Mixed	Yes	Yes	Yes	CVA
Type X	Mixed	Yes	Yes	Yes*	CVA

Yes*: including latent school factors

Issues of over-adjustments were found in some studies when controlling variables at classroom and school-level in the value-added models (McCaffrey et al., 2004; Ballou, et al., 2004). However, more recent research has argued the advantages of increasing the complexity of the school value-added models by specifying additional levels of variations, which can be either completely nested or follow a cross-classification pattern (De Fraine, Van Damme, Van Landeghem, Opdenakker, & Onghena, 2003; Goldstein et al., 2007; Leckie, 2009; Martínez, 2012; Rasbash, Leckie, & Pillinger, 2010; Timmermans, Snijders, & Bosker, 2013; Troncoso, Pampaka, & Olsen 2015).

In these studies, the levels of neighbourhoods, classrooms, primary and secondary schools, as well as local education authorities were specified (although not all levels simultaneously in all studies) and they were all found to be relevant sources of variation in pupils' test scores. However, these studies are based in the United Kingdom, the United States and the Netherlands, which are all developed countries whose education systems are embedded in radically different socio-economic contexts compared to the New Zealand education system.

This thesis focuses on contextualised value-added models (CVAs), as the primary objective is to develop models for more reliable intervention evaluation systems considering characteristics at all levels (i.e., student, teacher, and school) in the context of primary and secondary education in New Zealand. A reliable and frequently used

statistical method in contextualised value-added models is the multilevel model, which is discussed in detail later in Chapter 2.

1.3 Aims of the thesis and its contributions

The overarching research question is how to develop reliable and efficient value-added models to measure intervention effectiveness in the context of primary and secondary education in the special context of New Zealand's educational system.

In Section 1.1, I discussed the preferred choices of valued student outcomes in New Zealand as the New Zealand Curriculum levels for primary education and attainments or endorsements in qualifications, course, and standards for secondary educations. Often the preferred valued student outcomes are in the form of categorical variables (including binary outcomes). Moreover, there is a lack of uniform valued student outcome in New Zealand with respect to both cross-sectional and longitudinal educational research. Cross-sectionally, primary schools can choose formal assessments of their preference to assess student learning progress in reading, writing and mathematics. Longitudinally, as discussed, valued student outcome varies across junior primary schools, senior primary schools and secondary schools. It leads to the challenge of non-uniform outcome measures and the necessity of using outcome variables in the form of categorical data (where the binary outcome is a special case of categorical data with only two categories).

While very obvious and perhaps exaggerated in the New Zealand system the situation of needing to use categorical data to determine effectiveness is not unique. Other jurisdictions face this issue. For example in the United States, the value-added model often applies to high school graduation rates where the valued student outcome is binary (either pass or not pass) (Finch & Cassady, 2014).

In Section 1.2, I discussed the role of multilevel models as the preferred statistical method in realising the practical theory of value-added models in determining intervention effectiveness. A complex data structure such as the case in secondary schools in New Zealand leads to challenges in multilevel models in particular when the outcome variables are categorical (or binary).

As a result, the focus of this thesis is to build multilevel models for categorical outcome variables, to enable robust evaluations of large-scale intervention studies for improving primary and secondary education achievements in New Zealand. These datasets and their implications for analysis and interpretation will be illustrated by the use of two case

studies: one from a sustainability study based in primary schools; and another from a large-scale, longitudinal research and development programme based in secondary schools.

This study, thus, aims to fill in some of the limitations in previous work and contribute to school effectiveness research, around the issues of fitting standardised or norm-referenced outcome in a continuous outcome model, dichotomisation of the continuous valued student outcomes in multilevel models and fitting group-specific random effects at various levels to fit complex data structure. There is a need for robust modelling, judged as models which are both reliable and accurate. It will be accomplished by answering the following research questions in relation to the aim of building robust multilevel models for categorical outcome variables, with practical implications for educational research in New Zealand and abroad:

1. How to choose between valued student outcomes for your value-added models?
   (a) How to optimise reliability when fitting categorical outcome variables in a continuous outcome model?
   (b) How to reliably use dichotomised continuous outcome variables in a binary outcome model?

   In Chapter 3, we will compare between a continuous outcome model, a categorical outcome model and binary outcome models using STAR data from a primary school case study.
2. How to specify the model to achieve a good balance in the trade-off between model complexity and model validity?
   (a) How many underlying random effects should be fitted in a multilevel model with categorical outcome in order to reliably and effectively determine individual- and group-specific effects.

   In Chapter 4, we will compare between multilevel models with varying specifications of random effects at student-, subject-, and school-levels using an example from NCEA data at the secondary school level.

This thesis has both practical and theoretical contributions to national and international education research. In practice, it serves an excellent example of how to evaluate educational programs using categorical outcome variables, particularly relevant in the New Zealand context in both secondary- and primary-school settings. It addresses challenges of non-uniform outcome variables in cross-sectional and longitudinal educational research in New Zealand. This practical contribution also extends to

international research such as in the US (high school graduation) and other countries using criterion-based tools or longitudinal studies that extend to tertiary education.

It also makes a theoretical contribution to the development of a more refined approach in intervention evaluation, which can be applied to school and teacher evaluation in other countries. Moreover, its applications are beyond the field of educational studies, with extensions to social science, biostatistics, medical study, clinical trial study, and many more that attempt to measure effectiveness in data with the multilevel structure taking the form of categorical variables. In this respect, it adds to the fields of categorical multilevel modelling, in particular as an extension of earlier works by Gelman and Hills (2007), Sullivan, et al., (1999), and Woltman, Feldstain, MacKay, and Rocchi (2012), as well as the broader value-added models in school or intervention evaluation systems (Lai, et al., 2014; Lai, et al., 2018; Jesson, et al., 2018a; Jesson, et al., 2018b; Cervini, 2009; Timmermans, et al., 2011; Timmermans, et al., 2013).

1.4 Organisations of the thesis

As argued in the introduction, most intervention and evaluation research in an educational context will use multilevel modelling as its primary analysis tool. It is due to the 'nested' nature of students within classrooms, within schools, and within regions. However, there are practical difficulties with the outcome variables when these analyses are dichotomous (binary) or categorical. The remainder of this thesis will discuss these issues, including a variety of appropriate modelling methods for analysing these data, illustrated through case studies of primary- and secondary-based research projects.

In Chapter 2, I provide extended review and discussion on the multilevel model (MLM). In Section 2.1, I describe MLM in mathematics using an example of a two-level model. Model assumptions on linearity, normality, homoscedasticity, independence, and overdispersion as well as their diagnostic tools are discussed in Section 2.2. Section 2.3 summarises the statistical testing applied in MLM including testing of regression coefficients, testing of variance components and the likelihood ratio tests for model comparisons. A general framework of statistical procedural is given in Section 2.4. In Section 2.5, the statistical power of multilevel models is discussed briefly. Section 2.6 describes the estimation methods and statistical software and packages used in this thesis.

In Chapter 3, the first research question defined in the previous section is addressed through a case study of an intervention in primary schools with three different outcome variables (continuous, categorical, and binary). The background of the case study

including its research aim and design, participants and measures are described in Section 3.1. Three different outcome models are defined in Section 3.2 and compared in 3.3 for their model reliability and accuracy. Section 3.4 discusses the model reliability of each outcome model and practical implications of the finding.

In Chapter 4, the second research question defined in the previous section is addressed through a case study of an intervention in secondary schools using binary outcome variables. Section 4.1 describes the research aim and design as well as the participants and measures. In Section 4.2, four model specifications of random effects (student-, subject-, and school-specific random effects), are defined and compared in terms of model reliability and accuracy in Section 4.3. Section 4.4 discusses the trade-off between model complexity and model robustness.

In Chapters 5, I conclude my answers to the two research questions of this thesis.

2 STATISTICAL MODELS

This chapter describes the multilevel model (MLM). It serves two purposes: one is as an extension to the literature review discussed in subsections 1.2.4 and 1.2.5; the other is to introduce model specification, necessary for understanding the following two chapters, and in answering the research questions. Multilevel modelling is a well-established area of statistical theory, and much of this chapter follows instructions from the existing literature (Gelman & Hill, 2007; Bryk & Raudenbush, 1998; Hox, 2010; Agresti, 2013; Goldstein, 1986, 1995, 1997; Austin & Merlo, 2017).

Section 2.1 describes the multilevel model specification mathematically using an example of a two-level model. Model assumptions on linearity, normality, homoscedasticity and independence are discussed in Section 2.2. In Section 2.3, a non-technical introduction to the statistical testing of multilevel models is given. A four-step procedure of multilevel modelling is given in Section 2.4. Statistical powers and estimations are briefly discussed in Section 2.5 and 2.6 respectively.

2.1 Model specifications

In this section, I provide a formal mathematical description of a general multilevel model (Bryk and Raudenbush, 1988), which is the foundation of the model specifications and developments described later in Chapter 3 and 3. However, a complete understanding of the mathematics is not necessary to understand the concepts of multilevel models as means of value-added models to estimate intervention success.

Let us consider in this chapter a simple case where the outcome variable is a student's achievement in reading comprehension, and the covariates are students' prior

achievement in reading and teachers' pedagogical content knowledge (PCK which is the integration of subject expertise and teaching skills (Lai, et al., 2018)). Data follows a two-level hierarchical structure where students (level-1) are nested within classrooms (level-2 groups). Students' reading achievements (level-1 units) are conceptualised as individual observations and classrooms (level-2 units) are termed as groups.

Two models at different levels are developed simultaneously in order to investigate the relationship within a given level, as well as the relationship across levels. Level-1 models reflect the relationship between individual observations (i.e. students' reading achievements) and level-2 model study how the relationship between individual observations varies between groups (i.e., classroom variations) (Hofmann, 1997, Woltman, et al., 2012, Sullivan, et al., 1999).

2.1.1 Level-1 model: the relationship between individual observations

A level-1 model can be described as follow,

$$E(Y_{ij}) = \mu = g^{-1}(\eta_{ij}) = g^{-1}(\beta_{0j} + \beta_{1j}X_{ij} + \epsilon_{ij}) \qquad (2.1)$$

where:

- $Y_{ij} :=$ outcome variable i.e., students' reading achievements measured for student i from classroom j,
- $E(\cdot) :=$ μ, expectation or mean value of the distribution of students' reading achievements,
- $\eta :=$ linear predictor,
- $g(\cdot) :=$ link function which describes the relationship between the expectation μ and the linear predictor η,
- $X_{ij} :=$ value on the level-1 covariate i.e., students' prior achievements in reading,
- $\beta_{0j} =$ intercept for classroom j,
- $\beta_{1j} =$ regression coefficient associated with X_{ij} for the classroom j, and
- $\epsilon_{ij} =$ random error associated with student i from classroom j.

The level-1 model consists of three components which are established by three equations in the formulae above: a random variable with a given distribution with unknown mean, μ; a linear predictor, η; and a link function, g.

2.1.1.1 Random variable

The outcome variable, students' reading achievements, in this case, is considered a random variable following a given distribution with unknown mean μ. The expectation of a random variable is its mean, $E(Y_{ij}) = \mu$. It is an important concept to understand that one of the interests of a multilevel model is the unknown mean not the average of a sampled data.

2.1.1.2 Linear predictor

The linear predictor η, which is used to predict the outcome (i.e., students' reading achievements), integrates information about the covariates (e.g., students' prior achievements) through linear combinations of unknown coefficients e.g., β s. Each coefficient indicates the relative effect of a covariate on the outcome. This can be written in mathematics as $\eta_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \epsilon_{ij}$. For any function f with a valid inverse f^{-1}, the relationship holds where $f^{-1}(\eta_{ij}) = f^{-1}(\beta_{0j} + \beta_{1j}X_{ij} + \epsilon_{ij})$.

2.1.1.3 Link function

The link function g, describes the relationship between the linear predictor of covariates and the unknown mean of the random variable (i.e., students' reading achievements). In mathematics, such relationship reads $\mu = g^{-1}(\eta_{ij})$. It links the sample data to the unknown distribution so that estimation of the unknown parameter μ is linked to the estimations of unknown coefficients β s. Two most frequently used link functions, particularly in educational research, are the identity function, $g(\mu) = \mu$ and the logit function, $g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$. The choice is often made so that the domain of the link function is matched to the range of the unknown mean, μ.

When the student reading achievement is measured as a continuous variable (e.g., PAT Reading), the identity function is used where the mean is assumed to follow a normal distribution with a mean of μ and variance of σ^2. A critical assumption of any level-1 models with continuous outcome variables is that their level-1 errors ϵ_{ij} follows a standardised normal distribution with a mean of 0 and a variance of σ^2.

In cases where the student reading achievement is in the form of a binary outcome (e.g., below or at/above expected curriculum level) or a categorical outcome (e.g., reading curriculum level), the logit link function is used where the underlying distribution can be either Bernoulli, binomial, categorical, or multinomial. For a binary response, the variance of a binomial distribution is completed determined by the mean (as the binomial

variance is a function of the mean of the binomial distribution). When logit link function is used, the level-1 variance component σ^2 is not directly estimated as the variance of a binary random variable is determined by a function of its mean. Instead it is normalised at a constant value of $\frac{\pi^2}{3}$ and higher level variance and estimations of coefficients are all rescaled according to the normalisation.

2.1.2 Level-2 model: the relationship between groups

The level-1 model explains the individual regression for each group. The second level regression (level-2 model) tries to explain variation in regression coefficients (β_{0j} and β_{1j}). It describes the variability across multiple groups. Consider a level-2 model with both level-1 regression coefficients as outcome variables: an *intercepts and slopes as outcomes model*,

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + \nu_{0j} \quad (2.2)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + \nu_{1j} \quad (2.3)$$

where:

- β_{0j} = intercept for the j^{th} group,
- β_{1j} = regression coefficient associated with X_{ij} (slope) for the j^{th} group,
- $W_j :=$ the value on the level-2 covariate (i.e., teachers' pedagogical content knowledge),
- γ_{00} = overall mean intercept adjusted for W_j,
- γ_{10} = overall mean slope adjusted for W_j,
- γ_{01} = regression coefficient associated with W_j with respect to level-1 intercept,
- γ_{11} = regression coefficient associated with W_j with respect to level-1 slope,
- ν_{0j} = random effects of the j^{th} group adjusted for W_j on the intercept, and
- ν_{1j} = random effects of the j^{th} group adjusted for W_j on the slope.

The assumptions of level-2 models can be summarised as follows (Raudenbush & Bryk, 2002; Sullivan et al., 1999; Woltman et al. 2012):

- β_{0j} and β_{1j} jointly follow a multivariate normal distribution with the means and a covariance structure as,

$$\mathrm{E}(\beta_{0j}) = \gamma_{00}; \qquad \mathrm{E}(\beta_{1j}) = \gamma_{10}$$

$$\mathrm{var}(\beta_{0j}) = \tau_{00}^2; \qquad \mathrm{var}(\beta_{1j}) = \tau_{11}^2$$

$$\text{cov}(\beta_{0j}, \beta_{1j}) = \tau_{01}^2$$

- ν_{0j} and ν_{1j} jointly follow a multivariate normal distribution with the means and a covariance structure as,

$$\text{E}(\nu_{0j}) = 0; \qquad \text{E}(\nu_{1j}) = 0$$

$$\text{var}(\nu_{0j}) = \tau_{00}^2; \qquad \text{var}(\nu_{1j}) = \tau_{11}^2$$

$$\text{cov}(\nu_{0j}, \nu_{1j}) = \tau_{01}^2$$

$$\text{cov}(\nu_{0j}, \epsilon_{ij}) = \text{cov}(\nu_{1j}, \epsilon_{ij}) = 0$$

Multilevel models are differentiated from GLMs by fitting random effects (ν_{0j} and ν_{0j}) in addition to fixed effects ($\gamma_{00}, \gamma_{10}, \gamma_{01},$ and γ_{11}). However, model specifications in practice depend on the underlying pattern of variance in the level-1 intercepts and slopes (Hofmann, 1997). Intraclass correlation coefficient (ICC) defined as $\frac{\tau_{00}^2}{\tau_{00}^2+\sigma^2}$ is used to estimates how much variation in the outcome variable presents between level-2 groups. Alternative model specifications of the two-level model are well summarised in literature (Raudenbush & Bryk, 2002; Sullivan et al., 1999; Woltman et al. 2012). For detailed discussions about more complex model specifications of higher-levels, please see Raudenbush and Bryk (2002) and Bell, Ene, Smiley, and Schoeneberger (2013).

It is often useful to aggregate the level-1 model and level-2 model in a combined model (see Equation 2.4) by substituting Equations 2.2 and 2.3 into Equation 2.1 to allow for the classification of variables and coefficients in terms of the level of hierarchy they affect (Woltman et al. 2012).

$$E(Y_{ij}) = g^{-1}(\gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij} + \gamma_{11}W_jX_{ij} + \nu_{0j} + \nu_{1j}X_{ij} + \epsilon_{ij}) \qquad (2.4)$$

It is apparent from the combined model specification that errors are composite in MLMs ($\nu_{0j} + \nu_{1j}X_{ij} + \epsilon_{ij}$). This means errors are no longer assumed to be independent across the level-1 units. The two random effects ν_{0j} and ν_{0j} show that there is dependency among observations nested within each group. Furthermore, the assumptions of their

variance-covariance structure allow for a heterogeneous variance of the error terms (Sullivan et al. 1999).

2.1.3 More about naming conventions

At this stage, it may be useful to consider some other naming conventions that I use in this thesis. For example, throughout this thesis, I use the term 'outcome variables' instead of dependent variables or response variables, which are more common terms in the statistical literature. I prefer to use the term 'outcome variable' for three key reasons. Firstly, the term emphasises the student *outcomes*, which are central in the context of educational research. Secondly, it reduces or downplays the implied associations in "dependent variables" in that relationships are given before models were fit between "dependent variable" and "independent variables", e.g., the relationship between students' reading achievements and teachers' pedagogical content knowledge. Thirdly, one purpose of educational research is to detect potential *associations* between outcomes and other variables. Most studies within the field of education are observational or quasi-experimental studies, and causal inference should be made with caution. "Response variable" therefore is not a preferred term.

Furthermore, I prefer to use the term "covariates" in comparison to "independent variables" or "predictors". In addition to the reasons outlined above, and as discussed in the model specifications, not all effects are independent. There are dependencies in multilevel data. Hence I would avoid at all cost to use the term "independent variables". The term "predictors" implies the intent of applying model estimates for predictions of future events, which is not the goal in most educational research.

2.2 Model assumptions

A thorough understanding of the underlying model assumptions is essential in statistical modelling. There are in general five key assumptions to consider for GLMs: linearity, normality, homoscedasticity, independence, and overdispersion. However, there are disagreements in the literature on their necessities and the degrees of satisfaction to be met. (McCullough & Nelder, 1989; Breslow, 1996; Cameron & Trivedi, 1998; Dobson, 2002; Hoffman, 2004). Dobson (2002) argued that when model assumptions are met, residuals should independently and identically follow an approximately standardised normal distribution.

Diagnostics approaches of model assumptions include informal visualisation of residuals and formal hypothesis tests. There is at least one formal hypothesis test available for each graphical plot of residuals (Seber, 1980). However, I used mostly graphical analysis in this thesis as suggested by the literature that visual method provides more information such as detecting violations, indicating ways they can be corrected, as well as a sense of the potential impact of such violation (Seber, 1980; Cameron & Trivedi, 1998; Dobson, 2002; Hoffman, 2004).

Multilevel model as a generalised linear mixed model (GLMM) is an extension to the generalised linear model (GLM), which allows for correlations between observations within groups. MLMs have the same assumptions as GLMs (e.g., ANOVA, linear regression), with few modifications with respect to the nature of the clustered data structure. This section outlines a general framework of the underlying assumptions in GLMMs and provides a set of diagnostic assessments that can be applied to check if any model assumptions are violated. This framework and set of diagnostic procedures are used in both case studies analysed in this thesis (see later subsections 3.3.5 and 4.3.3).

2.2.1 Linearity

Linear relationships are not assumed between the outcome variable (e.g., students' reading achievements) and covariates. However, linearity is required for the relationship between covariates and the linear predictor. As described in subsection 2.1.1.3, in applications of identity function as the link function, the linear relationship of covariates to the mean outcome is established. However, in cases where logit function is applied, linearity is assumed only between covariates and the linear predictor. Moreover, the nonlinear transformation of covariates (such as the use of a power term) is allowed.

A simple but effective way to test linearity is to plot the model residuals (i.e., the difference between the observed data and model estimates) against fitted values (or each covariate). If the pattern looks like random variables scattered around zero, linearity is assumed to be met. Otherwise, a higher-order term of the covariate may need to be fitted to account for any non-linear patterns remained in the residuals.

2.2.2 Normality

The assumption of normality is not required for the outcome variables and hence the error terms at the lowest level. Different link functions can be applied to distributions such as binomial, categorical and logistic. However, error terms at higher levels are assumed to

jointly follow a multivariate normal distribution with the means and a covariance structure described in subsection 2.1.2.

Under multilevel linear models (i.e. using identity link function), the level-1 residuals are assumed to be normally distributed (and independent and homoscedastic) as for GLMs. Violations of normality assumptions in multilevel linear models will result in biased estimated standard errors of model estimations and statistical inference (Raudenbush & Bryk, 2002).

Graphically, the normal Q-Q plot of the residuals can be used to examine how closely the plot follows a diagonal line (Hox, 2002). Alternatively, the histograms or box-and-whisker plot of residuals can help assess deviations from normality. Statistical tests for normality such as Shapiro-Wilk or Kolmogorov-Smirnov test can also be applied as supplements to the graphical examination. As for level-2 residuals, summary statistics such as skewness and kurtosis can be generated and reviewed.

In this thesis, the normality assumption is checked using normal Q-Q plot of the residuals. However, normality is not considered a key criterion for valid model fit and accurate model estimations as suggested by Gelman and Hill (2007), as they argued that assumption of normality does not affect model estimations in multilevel models.

2.2.3 Homoscedasticity

The assumption of the equality of population variance (e.g., homoscedasticity or homogeneity of variance) does not need to be satisfied as multilevel models can account for this by modelling different specifications of the variance-covariance matrix. However, the assumption of equal variance should be checked after model fitting because when homoscedasticity presents in model residuals, the standard errors of the estimated coefficients are biased, which lead to incorrect significance tests and statistical inferences (Hoffman, 2004).

Homoscedasticity can be checked graphically using a scatter plot of the residual and fitted values. It provides a quick and simple way to detect any violations where a pattern of random scatters indicates satisfaction of the homoscedasticity assumption. Alternatively, hypothesis tests such as ANOVA can also be used to determine if the variance the residuals is equal across observations under the null hypothesis that variances are all equal. Greene (2000) described several applications of formal statistical tests for distribution assumptions.

In this thesis, I will use a scatter plot of the residual and fitted values to visually examine homoscedasticity. Hypothesis testing will only be used if a visual test fails.

2.2.4 Independence

GLM assumes independence where outcomes are identically and independently distributed random sample from the population and covariates are independent of each other. As an extension to allows for correlations between observations within groups, the multilevel model is specifically used when the assumption of independence is violated.

However, MLMs do assume that individual-specific random effects (level-1 residual errors, ϵ_{ij}) are independent of the covariates and group-specific random effects (level-2 residuals errors). Also, MLMs assume that residual errors at the highest orders are uncorrelated.

Residual plots described earlier for testing of linearity and homoscedasticity can be used to check for independence assumptions where patterns of random scatters indicates sufficiency.

2.2.5 Overdispersion

The assumption that the variance is equal to the mean is restrictive for models using binary or count outcome. Variances of the underlying distributions (e.g., binomial, Poisson) of such outcome variables equal to their means (Van Hoef & Boveng, 2007).

However, overdispersion occurs in those data, particularly in binary and count data, where observed variability is far greater than the expected. Models such as the quasi-Poisson (where over-dispersion is estimated) and negative binomial (where the variance is assumed to be greater than the mean) models should be used for over-dispersed data (Hoffman, 2004; Van Hoef & Boveng, 2007).

Overdispersion can also be examined through non-parametric dispersion test via standard deviations of the residual errors (Hartig, 2019).

In this thesis, I will use the combination of the histogram of the residual standard deviations and the non-parametric dispersion test to determine overdispersion in the model.

2.3 Statistical testing

In this subsection, I provide a non-technical introduction to the statistical testing of multilevel models. It is critical to understand the function of the statistical tests and be able to correctly interpret its results so that models can be rightfully compared during the model development and selection process and findings of the final model in relation to the research questions can be rigorously explained and discussed.

There are three types of statistical testing employed in multilevel models: (a) the testing of regression coefficients (i.e., fixed effects); (b) the testing of variance-covariance components (i.e., random effects) and (c) comparisons between different multilevel models.

2.3.1 Testing of regression coefficients

In multilevel models, coefficients are assumed to jointly follow a multivariate normal distribution with the means and a covariance structure described in subsection 2.1.2. It follows from the assumption about multivariate normal residuals.

Statistical significance of regression coefficients in multilevel models is often determined by t-statistics using the Student's T-test. The t-statistics is the ratio of the estimated coefficient and its standard error. The underlying distribution of the t-statistics is the Student's t-distribution with a given degree of freedom. Alternatively, T-test can be replaced by the Z-test as the t-distribution approaches the standardised normal distribution as the sample size increases. The z-statistics is identical to the t-statistics. Another alternative is the Wald statistic (z-statistics squared), which asymptotically follows the chi-square distribution.

Significances of regression coefficients can also be assessed by the likelihood ratio test, particularly in multilevel models with the logit link function. The likelihood (L) of a statistical model is a function of the coefficients given specific observed data that describes the plausibility of the estimated coefficients. High likelihood implies more plausible (i.e., better) model fit. The likelihood ratio test compares the goodness-of-fit of the coefficient model with that of the null model (e.g., model without the covariates of interest). The likelihood ratio test requires nested models where one model is a subset of the other using the same data. The goodness-of-fit statistics used in the likelihood ratio test is deviance, which is defined as $-2\ln(L)$. Although the underlying distribution of deviance is unknown, the difference between deviances (or likelihood ratios) of two multilevel models is assumed to follow a chi-squared distribution with degrees of freedom

set by the difference in the number of covariates fitted in the two models. Coefficients are considered statistically significant if the deviance of the coefficient model is significantly lower than the null model one.

2.3.2 Testing of variance components

Testing of variance components is more complicated as their underlying distributions are skewed and not normal, especially when the size of groups is small. The likelihood ratio test is also recommended as a way of testing the statistical significance of variance components. Similar to the testing of regression coefficients, by comparing the deviance of two multilevel models, one with the random effect of interest and the other without, the variance component is significant if the deviance of the model with the random effect is significantly lower.

2.3.3 Model comparisons

As discussed in the previous two subsections, the likelihood ratio test can be applied to compare two nested models. In general, it is assumed that the lower the deviance, the better the model.

When the models are not nested and differ in both fixed and random effects, the Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) can be used to determine the model fits. Given a sample data Ω and let n be the number of observations in the sample and k be the number of estimated coefficients in the model, the AIC of a multilevel model is defined as $2k - 2\ln(\max_{\Omega} L)$ and the BIC is $\ln(n)\,k - 2\ln(\max_{\Omega} L)$. Both the AIC and the BIC penalises the model complexity where overfitting is discouraged. Given a set of multilevel models using the same data, the model with the lowest AIC and/or BIC is preferred.

2.4 Statistical procedure

Reliable model estimation depends on not only an appropriate model specification, a thorough check of model assumptions but also an excellent statistical procedure. While the model building process when estimating multilevel models (regardless of their link functions) vary between researchers and guidelines from difference sources, they share three elements in common: (a) the aim is to estimate the simplest models that best describe the data; (b) always build an empty model (i.e., unconditional model without covariates)

first so that its ICC can be calculated; and (c) stepwise build more complex models while testing for significance using likelihood ratio tests after each model estimation.

I propose in this subsection a four-step statistical procedure for multilevel models as follows:

i. Step 1: Build an empty model, which includes only random effects without any covariates. Compute the ICC of the empty model. If the ICC is negligible (i.e., not different from zero), then a GLM such as multiple linear regression can be used instead. Otherwise, move to Step 2.

ii. Step 2: Add level-1 covariates, as well as their interactions. Using likelihood ratio tests to determine the statistical significance of each covariate and each interaction terms. Keep or remove covariates and interactions based on the significance of likelihood ratio tests. Compare models with and without theoretically relevant random slopes and you may consider removing the random slope terms if they do not improve the fit based on likelihood ratio tests.

iii. Step 3: Repeat Step 2 adding level-2 covariates and their intra-level interactions.

iv. Step 4: Repeat Step 2 adding the cross-level interactions.

Finally, the AIC and BIC can be used to determine the simplest models that best describe the data as they penalise the number of coefficients estimated in the model.

2.5 Statistical power

Statistical power for multilevel models varies depending on (a) effect size and intraclass correlations; (b) the number of groups and the number of individual observations per group, (c) whether it is for fixed effects or random effects and (d) whether it is for level-1 or level-2 covariates. Maas and Hox (2005) provide an excellent review of the sufficient sample sizes for multilevel models.

Intuitively, sufficient power requires large sample sizes in multilevel models. As suggested in the literature, a sample size of 150 individual observations is sufficient to achieve 80% of power for level-1 effects, and group size of 20 is needed to detect cross-level interactions as estimations of level-2 standard errors can be biased when the number of groups is small. A sufficient number of groups is also required in general in order to achieve convergence in model estimations. Furthermore, Gelman and Hills (2007) offer calculations of sample sizes given a set of requirements such as achieving a specified standard error or probability of obtaining statistical significance.

2.6 Statistical Computation

Estimations of multilevel models are usually performed by maximum likelihood (ML) algorithm (Hox (2002, Chapter 3)). In models with the logistic link function, one common estimation algorithm is the Laplace estimation. Estimation theory is a complex field in statistics that is beyond the scope of this thesis. For more detailed explanations, please see Rabe-Hesketh and Skrondal (2012), Chapter 2.10-11, and in Raudenbush and Bryk (2002), Chapter 3.

In this thesis, models were estimated using `R` (Version 3.5.4, R Core Team, 2018) with the `lme4` package (Bates, Maechler, Bolker & Walker, 2015a, 2015b). Diagnostic tests are performed using the `DHARMa` packaged in R based on simulations of residual errors from each model fit (Hartig, 2019).

3 A PRIMARY SCHOOL CASE-STUDY

This chapter aims to address the challenge in multilevel modelling when there is an absence of a unique valued student outcome. It aims to answer the first research question: "How to choose between valued student outcomes for your value-added models?" through a case study of longitudinal primary school intervention.

The dataset used in this study was from a larger project led by the Woolf Fisher Research Centre that I was a part of as the lead statistician. The data for this thesis uses a portion of the data from this project to examine the relationship between pedagogical content knowledge generated collectively (CPCK) in a professional learning community (PLC), and student achievement in reading comprehension (Lai, et al., 2018). The background of the study and the methods were published in Lai, et al. (2018), and are paraphrased here so that the reader understands the context of the case study. The statistical modelling described here extends the study by comparing models using different valued student outcomes. In the published study, I developed a continuous valued outcome and modelled by a hierarchical model with two levels for the publication. In this chapter, two other valued student outcomes are fitted using the same model specifications described in the published study: one is a categorical data with nine ordered levels, and the other is a binary data. Also, unlike the publication, PCK here is treated as an intervention. This chapter aims to compare the results of model estimations and goodness-of-fit of these two models with the published one.

I describe the study background and discuss the dataset and transformations of the original student outcome measure in Section 3.1. In Section 3.2, I provide model specifications mathematically and describe its link function for each outcome model. I present and compare results on intraclass correlation coefficient, goodness-of-fit, variance explained, model estimations of fixed effects and diagnostic tests of model assumptions in Section 3.3. In Section 3.4, I summarise and discuss the key finding and its practical implications.

3.1 Background to the study

The case study is part of a longitudinal intervention study that took place in primary schools (n = 8) across two low socio-economic communities in New Zealand. The intervention logic from that paper is summarised here so that the reader understands the rationale for the model. PCK is the knowledge of a subject matter and how to teach it (Shulman, 1986). In a PLC, PCK is collective in nature, as knowledge is shared and distributed across the PLC (Wenger, 1998). The definition of collective PCK (CPCK) here is contextualised to data discussions where the aim is to analyse data and create or modify instructional practices to improve student learning (Lai, et al., 2018). "*In those data discussions, collective knowledge building occurs when participants become more specific about the instructional practices that will address an achievement problem e.g., describe more details about the practice and under what conditions it is best implemented"* (Lai, et al., 2018). Empirical studies suggest that increased PCK specificity is linked to improved achievement (e.g., Lai, et al., 2014). Linking PLCs with achievement assumes a chain of influence mediated through CPCK (Lai, et al., 2018).

The goal of the CPCK study was to examine the relationship between variations in CPCK and variations in achievement. Examining this relationship at this juncture could provide valuable information about the concept of CPCK, as well as point to potential relationships which others can use to develop causal models.

In the following subsections, I first describe the participants involved in the wider intervention study in subsection 3.1.1. I then describe the measures collected in the wider study and the reduced case study data for this thesis in subsection 3.1.2.

3.1.1 Participants

The original study includes 1082 students across five different year level cohorts (Year 4-5, 5-6, and 6-7, 7-8, and 8-9) from eight schools. Schools come from two clusters of

low socio-economic communities with high proportions of indigenous and ethnic minority students, and similar proportions of males and females ($\chi 2(1) = 0.26$, $p = 0.61$).

There were approximately 120 teachers and 23 school leaders participated in the study. About two-thirds of teachers had at least five years of teaching experience, and around 10% were beginning teachers (Lai, McNaughton, Amituanai-Toloa et al., 2009).

3.1.2 Measures

3.1.2.1 Valued student outcomes: reading achievement

STAR achievement data were collected at the beginning of two school years. It includes scaled scores (continuous), stanine scores (categorical) and reading curriculum levels (categorical) aligned with the framework set by the NZC, as well as student demographics. Student reading achievements at the first data collection are considered a measure of prior achievements. Descriptions of student characteristics (as a percentage of the total sample) are shown in Table 3.1.

As discussed in subsection 1.1.1, adjustments of scaled scores to a set of norms are preferred because students at different year levels are expected to follow different learning curves. There are three types of transformations available: (a) norm-adjusted scale scores as the difference of scaled scores to the year-specific norms (see Table 1.3); (b) stanine scores; and (c) a binary outcome indicating at/above or below (year-specific) expected curriculums.

Each transformation is straightforward, and each transformed outcome is valid and reliable. However, the use of norm-adjusted scale scores is more prevalent in educational studies (Lai, et al., 2018; Lai, et al., 2014; Jesson, et al., 2018a; Jesson, et al., 2018b).

Norm-adjusted scale score maintains the underlying distribution of the original scale score with the same underlying variance. Stanine, on the other hand, categorised the original scale score to nine bins where the underlying distribution is approximately normally distributed with a mean of five and a standard deviation of two. The binary indicator dichotomised the original scale score to a zero/one indicator which follows a Bernoulli or binomial distribution where only the mean is estimated as its mean determines the variance.

In this chapter, I compare models under the same model specifications for each transformed outcome: (a) the *continuous outcome model* (using the norm-adjusted scale

score); (b) the *categorical outcome model* (using the stanine score); and (c) the *binary outcome model* (using the binary indicator in relation to expected curriculum levels).

Table 3.1 Descriptive Summary of Student-level Covariates in the Primary School Case Study

Student Characteristics	Original (n = 1082)	Case study (n = 573)
Gender		
Male	48.9%	50.1%
Female	51.1%	49.9%
Ethnicity		
Maori	16.5%	15.5%
Pasifika	79.5%	79.6
Other	4.0%	4.9%
Year Cohort		
Year 4-5	23.49%	-
Year 5-6	29.0%	-
Year 6-7	11.1%	-
Year 7-8	27.4%	-
Year 8-9	8.5%	-
Prior Achievements (Norm-adjusted Scaled Scores)		
Mean	-15.9	-16.1
Standard Deviation	11.7	11.9

3.1.2.2 Observations of PLC

Eight PLC meetings (one per school) where data were discussed and were taped. The unit of analysis for the PLC meetings was a turn, which was each time a person spoke. CPCK-related turns were scored from 1-3. A higher score indicated that more specific instructional practices were discussed to solve an achievement problem. Non-CPCK turns were analysed thematically into Data Trend (discussing patterns in the data) and Other. Inter-rater reliability was 86% (Lai, et al., 2018). PCK scores were aggregated to form a CPCK score for each PLC.

3.1.2.3 Case study data for model comparison

The objective of this chapter is to study differences in model robustness in terms of model reliability and accuracy under the same model specifications using three different types of outcome variables (e.g., continuous, categorical, and binary). As a result, it is important to compare model fits and behaviours under an identical dataset. In this subsection, I describe a subset of the case study data that is used in this chapter (see Table 3.1).

In this case study, I include only students from Year cohort 4-5 and 5-6 because (a) students from higher year cohorts are considerably underperforming and require more refined model specifications which are not the focus of this thesis (Lai, et al., 2018); (b) the reduced dataset has enough sample size (n = 573) at the student-level for multilevel models; (c) student characteristics are similar to the whole dataset as shown in Table 3.1).

Three outcome variables are considered (i.e., norm-adjusted scale score, stanine score, and binary indicator relating to the expected curriculum level). The same model specification is applied in each outcome model including (a) two student-level covariates i.e., prior achievement and year cohort and (b) one school-level covariate i.e., combined proportions of turns involving CPCK level 1, 2, 3 and Data Trend (*CPCK+DT*).

In the following section, I describe the models and model specifications mathematically.

3.2 The models and model specifications

In this chapter, multilevel models with two hierarchical levels (i.e., students nested within schools) are used to examine (with respect to the research question of the original study) possible relationships between variations in CPCK and level of achievement at the end of the intervention, which is the second data collection point (Lai, et al., 2018).

With respect to the research question of this chapter, I built three multilevel models which differed only by outcome: (a) a continuous outcome model, where the valued student outcome is the norm-adjusted scale score; (b) a categorical outcome model, where the outcome is measured in stanine scores; and (c) a binary outcome model, where the outcome is a binary indicator of meeting or not meeting the expected curriculum level.

For each outcome model, the same model specifications and modelling procedurals were followed. Statistical procedural follows the steps described in Section 2.4 where an empty model is fitted first, followed by a baseline model with all relevant student-level covariates, and finally the full model with the inclusion of school-level covariates. As a result, for each outcome model, there are three fitted model specifications i.e., the empty model, the baseline model, and the full model (see Table 3.2). For ease of reading, in this thesis, I refer to the baseline model of the binary outcome model as the baseline binary outcome model and so on.

Table 3.2 Model Specifications of Each Empty, Baseline, and Full Model in the Primary School Case Study

Estimated Effects	Empty Model	Baseline Model	Full Model
Random Effects			
Student-specific	Estimated	Estimated	Estimated
School-specific	Estimated	Estimated	Estimated
Fixed Effects			
Prior Achievements	-	Estimated	Estimated
Year Cohort	-	Estimated	Estimated
CPCK+DT	-	-	Estimated

In what follows, I will first describe the model specifications which are the same for all three outcome models. Then I will describe the three outcome models (continuous, categorical, and binary) and their corresponding link functions.

3.2.1 Model specifications

The three model specifications are nested as shown in Table 3.2. The empty model defines the multilevel structure with only random effects. The baseline model adds the student-level covariates which explain variations between students within a school. The full model adds school-level covariates which explain variations of between-school effect, i.e., the intervention effect of *CPCK+DT*.

Due to its nesting structure, I present here only the full model specification as the baseline and empty model specifications can be obtained by removing covariates that are not estimated. Let Y_{ij} be reading achievement outcome for student i in school j and define η_{ij} as the linear predictor of the multilevel model (described in subsection 2.1.1.3), the *full* model is defined as:

$$\eta_{ij} = \beta_{0j} + \beta_1 Prior_{ij} + \beta_2 C_{ij} + \epsilon_{ij} \quad (3.1)$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01} CPCKDT_j + \nu_{0j} \quad (3.2)$$

where:

- β_{0j} = intercept for school j,
- $\beta_1,\ \beta_2$ = regression coefficient associated with each student-level covariate,
- ϵ_{ij} = random effects of student i in school j,
- $CPCKDT_j :=$ the proportion of CPCK+DT for school j,
- γ_{00} = overall mean intercept adjusted for the school-level covariate,
- γ_{01} = regression coefficient associated with school-level covariate with respect to student-level intercept,
- ν_{0j} = random effects of school j adjusted for school-level covariate on the intercept.

Student-specific effects ϵ_{ij} and school-specific effects ν_{0j} follows the normality assumption as described in subsection 2.1.2.

3.2.2 Link functions

Each outcome model is built under the same model specifications which differ only in the underlying distribution of the outcome variable: (a) the continuous outcome model assumes normal distribution; (b) the categorical outcome model also assumes (approximately) normal distribution; (b) the binary outcome model assumes a Bernoulli distribution.

As a result, the identity link function is applied in both the continuous and categorical outcome models where,

$$E(Y_{ij}) = \mu = g^{-1}(\eta_{ij}) \qquad (3.3)$$

On the other hand, in the binary outcome model, the logit link function is applied where,

$$\text{logit}\, \mathrm{E}(Y_{ij}) = \text{logit}\, \Pr(Y_{ij} = 1) = \log\left(\frac{p}{1-p}\right) = \eta_{ij} \qquad (3.4)$$

where p is the unknown probability of meeting the expected curriculum level and η_{ij} is the linear predictor of its log-odds.

As discussed in Section 2.2, model estimations from the binary outcome model using the logit link function with Bernoulli distribution are recalled so that the level-1 residual variance is constant at approximately 3.29. As a result, direct comparisons of fixed and random effects should be avoided between other model estimations.

I present and compare in the next section, model goodness-of-fits and estimations of each outcome model.

3.3 Results

The results are organised to enable comparisons across the three outcome models. For each outcome model, three model specifications were fitted: empty, baseline and full. Therefore, in total nine models were tested and reported on. I first discuss the intraclass correlation coefficients (ICC) of the empty model across three outcome models, which establish model reliability of the school-specific random effect in subsection 3.3.1. In subsection 3.3.2, I compare model goodness-of-fit between the empty, baseline, and full across three outcome models. Subsection 0 describes the variance explained by each full model with respect to each empty model. I present in subsection 3.3.4 model estimations of fixed effects. In subsection 3.3.5, I examine model assumptions of each full model across the three outcome models. For a complete set of model estimations, please see Appendix A to I.

3.3.1 The intraclass correlation coefficient (ICC)

As described in Section 2.4, the first step of modelling is to build an empty model so that ICC can be calculated to evaluate the strength of correlation within schools (Bliese, 1998;

Fleiss, 1986; James, Demaree & Wolf, 1985; LeBreton & Senter, 2008). Three empty models were fitted for each outcome model (continuous, categorical, and binary). Table 3.3 summarises the estimated variances of the random effects of each empty model at each level and the corresponding ICCs.

Table 3.3 Summary of Variances of Random Effects of Empty Models and ICCs in the Primary School Case Study

Variance of	Continuous Model	Categorical Model	Binary Model
Individual Random Effect (Student-level)	125.40	2.85	3.29
Group Random Effect (School-level)	11.06	0.30	0.16
ICC	8.10%	9.5%	4.75%

ICCs of all three models are above the suggested cut-off point at 4% for reliable estimations of school-level effects (Fleiss, 1986). Moreover, ICCs of continuous and categorical outcome models are both higher than 7.5%, a benchmark indicating excellent reliability of school-level means.

It is recommended that if the ICC is negligible (i.e., not different from zero), then the higher-order random effect can be removed and a generalised linear model such as the simple multiple linear regression can be used instead (Bliese, 1998; Fleiss, 1986; James, et al., 1985; LeBreton & Senter, 2008).

However, the ICC of each empty model suggests that school-specific random effect is relevant to model estimations of school effects. Therefore, results on ICCs are consistent across three outcome measures.

3.3.2 Goodness-of-fits

Following the suggested procedurals, three baseline models with two student-level covariates: cohort and prior achievements were then fitted. Then three full models were fitted with an update of the school-level covariate: *CPCK+DT*. In the following subsections, I compare the nested model specifications (e.g., empty to baseline to full) across three outcome models using goodness-of-fit measures (including AIC and

deviance) and likelihood ratio test statistics (such as the chi-squared statistics ($\chi 2$), degree of freedom (df) and p-value) to determine reliability of each full model.

3.3.2.1 Continuous outcome model

Summary of goodness-of-fit measures (including AIC and deviance) and likelihood ratio test statistics (such as the chi-squared statistics (χ^2), the degree of freedom (df) and p-value) are shown in Table 3.4. Note that the results of the likelihood ratio tests shown in the table are comparisons between successive fits, i.e., comparisons between the baseline and the empty model and comparison between the full and baseline model.

Table 3.4 Summary of Goodness-of-fit Measures and Likelihood Ratio Test Statistics of the Continuous Outcome Model

Continuous Outcome	Empty Model	Baseline Model	Full Model
AIC	4415.0	3753.3	3750.1
Deviance	4409.0	3743.3	3738.1
Likelihood Ratio Test			
χ^2	-	665.78	5.13
df	-	2	1
P-value	-	< .001	0.02

Reductions in deviance and AIC were found from the empty model to the full model as shown in Table 3.4. Significances of the likelihood ratio test also suggested that full model with the cohort, prior achievement, and the *CPCK+DT* best describes the sampled data using the continuous outcome variable (i.e., norm-adjusted star scale score).

3.3.2.2 Categorical outcome model

Similarly, Table 3.5 shows that the full model better explains variations in student reading achievement (measured as stanine) in comparison to the baseline and empty models (under the same model specifications of fixed and random effects as the continuous outcome model).

Table 3.5 Summary of Goodness-of-fit Measures and Likelihood Ratio Test Statistics of the Categorical Outcome Model

Categorical Outcome	Empty Model	Baseline Model	Full Model
AIC	2248.5	1632.1	1627.2
Deviance	2242.5	1622.1	1615.2
Likelihood Ratio Test			
χ^2	-	620.37	6.98
Degree of Freedom	-	2	1
P-value	-	< .001	< .001

3.3.2.3 Binary outcome model

As for the binary outcome model (see Table 3.6), the full model including the additional school-level effect was marginally better than the baseline model (p = 0.08). It means that the reduction in deviance is found but not statistically significant at 5%.

Table 3.6 Summary of Goodness-of-fit Measures and Likelihood Ratio Test Statistics of the Binary Outcome Model

Binary Outcome	Empty Model	Baseline Model	Full Model
AIC	501.3	296.4	295.3
Deviance	497.3	288.4	285.3
Likelihood Ratio Test			
χ^2	-	208.83	3.11
Degree of Freedom	-	2	1
P-value	-	< .001	0.08

In summary, based on the AICs and deviances and following the rule the smaller, the better, each full model is considered the better model in goodness-of-fit in comparison to each baseline model indicating reliably model fit on school-specific effects.

3.3.3 Variance explained

In this subsection, I present and compare (across three outcome models) two different statistics: marginal squared residual error and conditional squared residual error, which are calculated based on methods given by Nakagawa, Johnson, and Schielzeth. (2017). Marginal squared residual error (R_m^2) measures variance explained by the fixed effects and conditional squared residual error (R_c^2) represents variance explained by the full model (i.e., including both fixed and random effects) (Nakagawa & Schielzeth, 2013; Nakagawa, et al., 2017).

Table 3.7 Summary of Variances of Random Effects of Full Models in the Primary School Case Study

Variance of	Continuous Model	Categorical Model	Binary Model
Individual Random Effect (Student-level)	39.71	0.98	3.29
Group Random Effect (School-level)	0.77	0.01	0.00
ICC	1.9%	1%	0%

Variances of random effects of each full model (continuous, categorical, and binary) at each level (student- and school-level) are presented in Table 3.7. Compared with random effect variations of their corresponding empty models (see Table 3.3), variations have dropped at both levels across three full models. However, almost all the unexplained variance is due to unmeasured variances between students across three full models.

As discussed in subsection 2.1.1.3, when the logit link function is applied (as in the case of the full binary outcome model), school-level random effect variation is rescaled so that the variation of the student-level random effect is constant at $\frac{\pi^2}{3}$ (approximately 3.29). Variation of the school-level random effect of the full binary outcome model is estimated at 0 indicating a singular fit. This implies overfitting in the full binary outcome model indicating that the observed data may not support the school-specific effect. This leads to the consideration of removing such random effects so that a more parsimonious model is achieved.

However, as discussed in earlier subsections (e.g., subsection 1.2.4.2), statistical inference in relation to the research question should be considered before removing any random effects, in particularly in the case of a singular fit where removal of the random effect will not change the estimated school effects, but only improve its goodness-of-fits (such as the AIC and deviance).

Table 3.8 Marginal and Conditional Residual Squared Errors in the Primary School Case Study

Model Specification	Residual Error Squared	Continuous Model	Categorical Model	Binary Model
Baseline	Marginal	68.7%	66.6%	50.1%
	Conditional	70.2%	67.8%	
Full	Marginal	69.9%	67.8%	50.5%
	Conditional	70.5%	68.1%	

It is found that the full continuous outcome model explained 71% of the variations in its data, whereas the full categorical and binary outcome model explained 68% and 51% of their data variations respectively (see Table 3.8). We should be cautious in making direct comparisons between models using variance explained because of the rescaling and normalisation in model estimations in the full binary outcome model (Snijders & Bosker, 1999).

Variance explained by the full continuous outcome model and the full categorical outcome model indicate reliable model estimations, while the singular fit is found in the full binary outcome model.

3.3.4 Estimations of fixed effects

The objective of each full model with respect to the original research question is to examine value-added effects by improved CPCK and Data Trend discussions in PLC. In the following subsection, I compare model estimations of the three full models in directions and statistical significance. Comparisons in the size of effects between models are not discussed.

Table 3.9 Model Estimations of Fixed Effects of the Full Models in the Primary School Case Study

Covariate	Continuous Outcome	Categorical Outcome	Binary Outcome
Fixed Effects	Coef. (s.e.)		
Intercept	-5.50 (4.21)	4.63 (0.58)	-1.95 (1.59)
Cohort: Year 5-6	-2.01 (0.55)	-0.85 (0.09)	-0.60 (0.34)
Prior Achievement	0.82 (0.02)	0.12 (0.003)	0.23 (0.03)
CPCK+DT	17.92 (7.72)	3.01 (1.07)	5.24 (2.83)

Table 3.9 presents the estimated fixed effects of each full model specification. When comparing the estimates in Table 3.9, the direction of the relationships is consistent for all fitted covariates across the three outcome models. Take intercept, for example, a negative norm-referenced scale score, a stanine below 5, and a negative log-odds all indicate achievement below expectation.

The statistical significance of the relationships, on the other hand, is maintained between the full continuous outcome and full categorical outcome model, while the statistical significance of the estimated effect of the *CPCK+DT* and year cohort are marginal (i.e., significant at 10%, not 5%). It means that estimations of the full binary model are more conservative than the other two.

3.3.5 Assumption checking

As discussed extensively in Section 2.2, checking for model assumptions are critical in the multilevel analysis. In the previous subsections, I showed that all three full models have good reliability of model estimations, acceptable goodness-of-fit, and fair to explain the observed variations among different student outcomes. However, without checking for assumptions, the accuracy of model estimations remains uncertain. In the following subsection, I provide a set of diagnostic tests to examine model assumptions for each full model across the three outcome models (continuous, categorical, and binary).

A set of residual plots (see the right panel on Figure 2, Figure 4, and Figure 6) is presented in the following subsection in order to examine model assumptions of linearity, homoscedasticity, and independence. The scatter plot of the standardised residuals and fitted values reasonably behaved across three full models where general patterns of

random scatter around zero emerged (see the right panel on Figure 2, Figure 4, and Figure 6). Equal and constant variations among observations are also found. These residual plots also validate linearity assumptions.

Q-Q plots of residuals (see the left panel in Figure 2, Figure 4, and Figure 6) show no deviations from the standardised normal distribution indicating satisfactory of normality assumptions across three full models.

Histograms of the standard deviations of the residuals are presented (at the centre) in Figure 3, Figure 5, and Figure 7 to visually inspect patterns of overdispersion. No overdispersion is found across the three full models.

Histograms of simulated residuals are also shown (on the far right) in Figure 3, Figure 5, and Figure 7 where outliers are highlighted in red and are also tested based on the exact binary tests. No influential outlier is found across the three full models.

Overall, I consider that assumptions of linearity, normality, homoscedasticity, independence, and overdispersion are met for all full models across three outcome models i.e., the continuous outcome model, the categorical outcome model, and the binary outcome model.

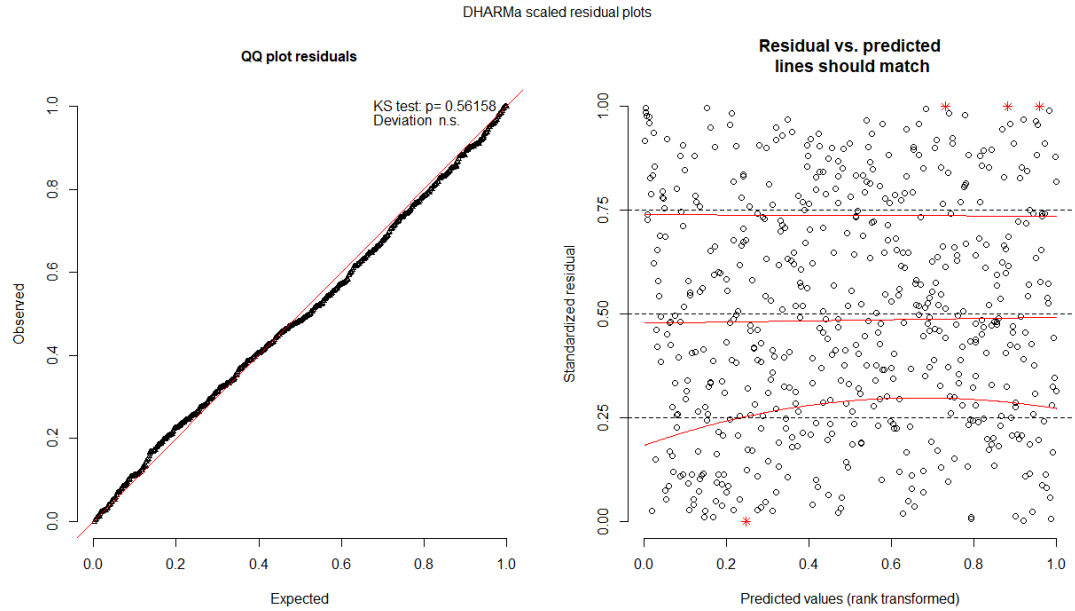

Figure 2 Residual versus fitted values and the residual normal Q-Q plot of the full continuous outcome model.

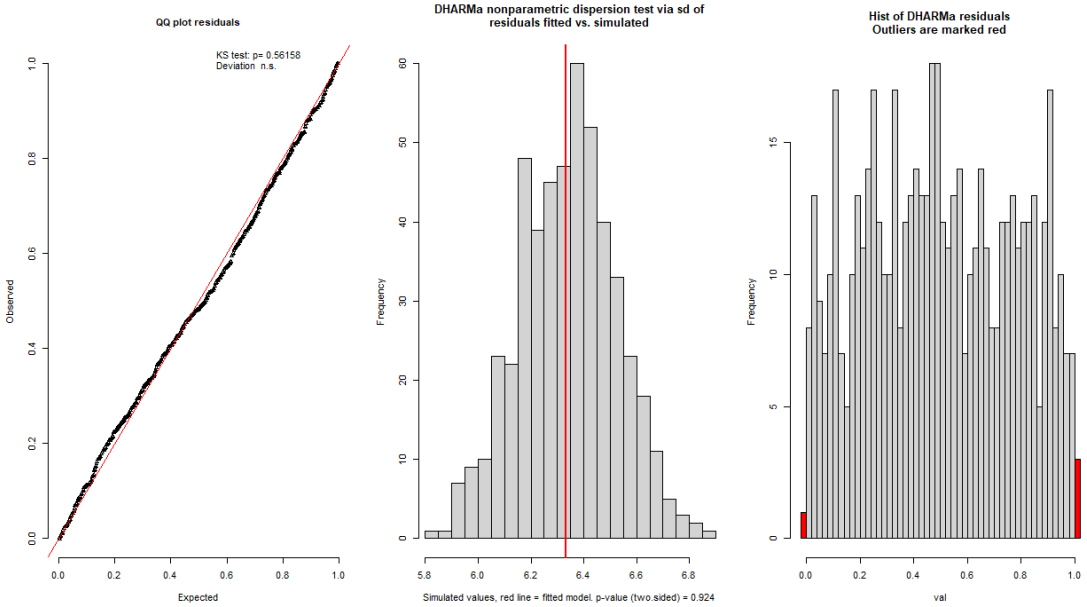

Figure 3 Tests on normality, overdispersion, and outliers of the full continuous outcome model.

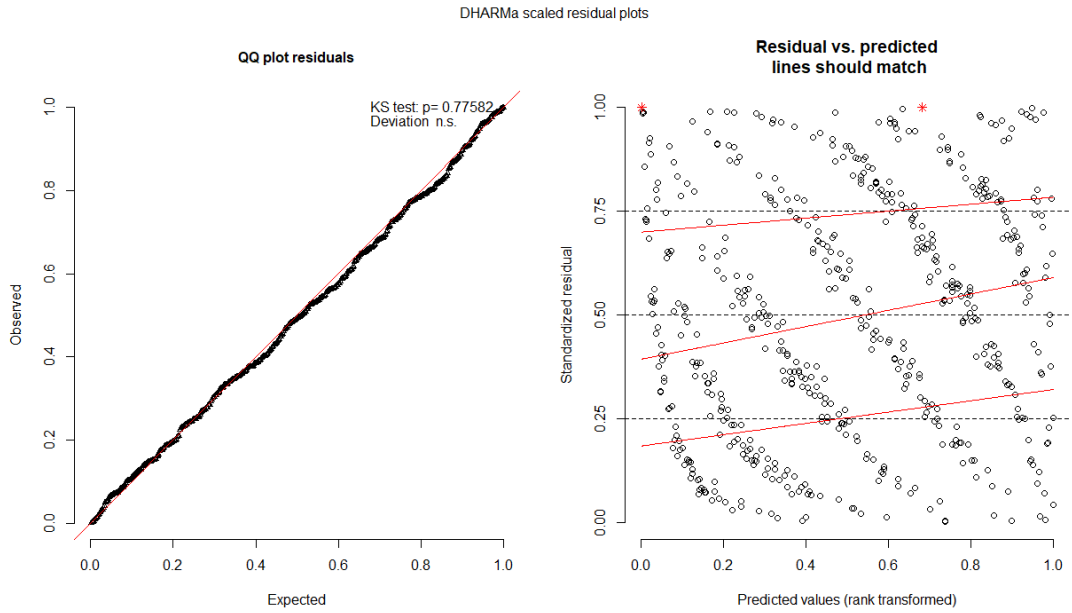

Figure 4 Residual versus fitted values and the residual normal Q-Q plot of the full categorical outcome model.

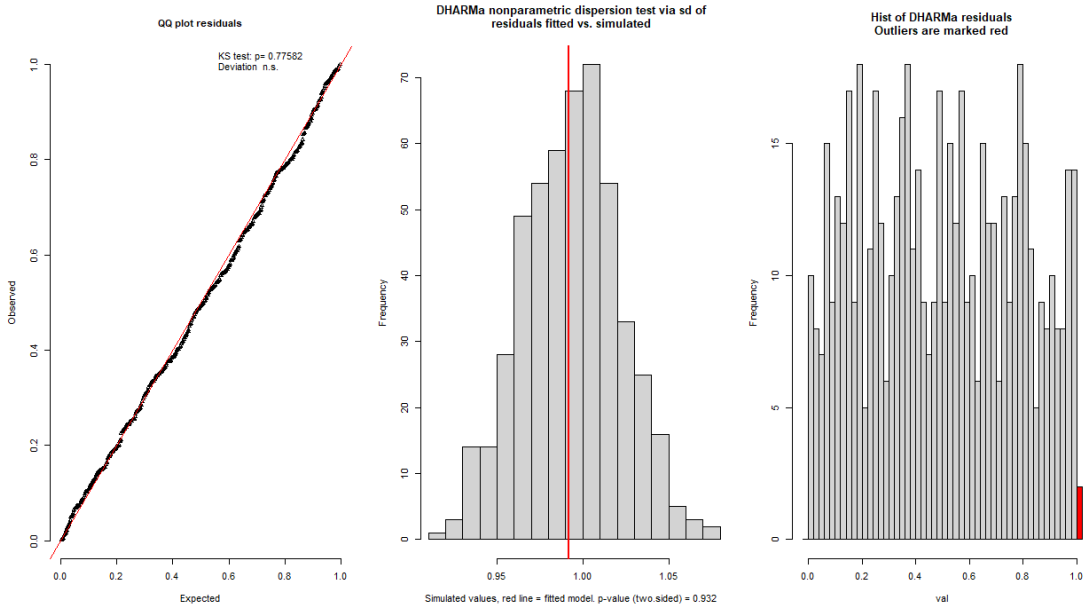

Figure 5 Tests on normality, overdispersion, and outliers of the full categorical outcome model.

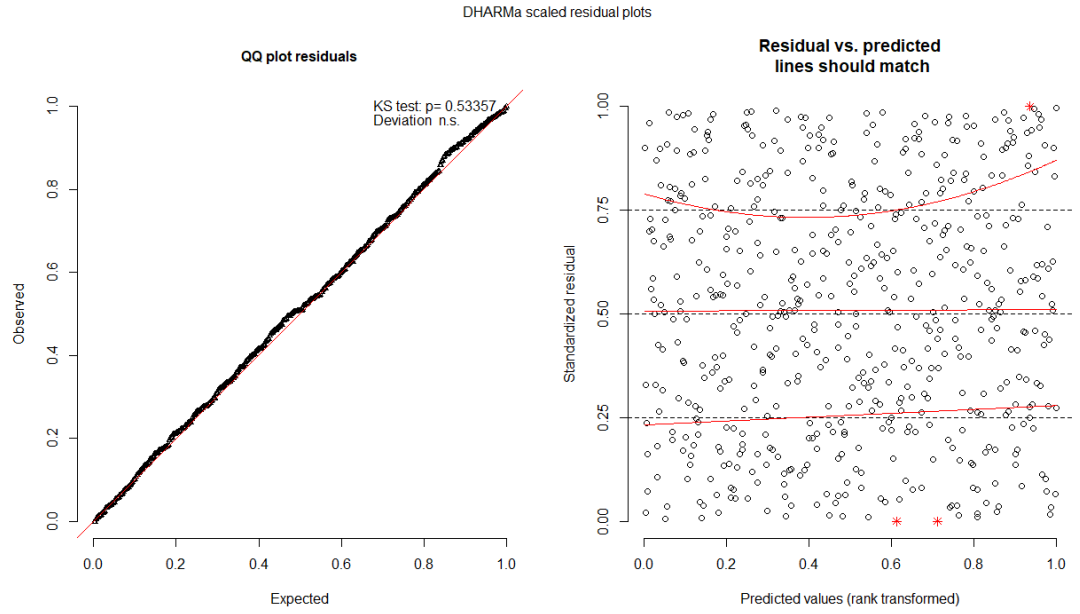

Figure 6 Residual versus fitted values and the residual normal Q-Q plot of the full binary outcome model.

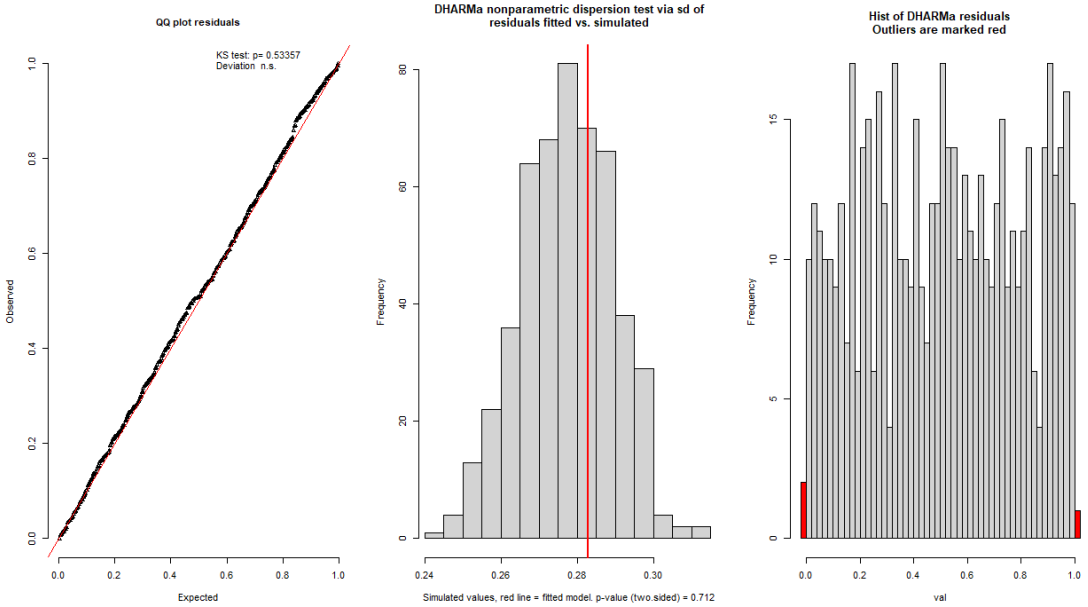

Figure 7 Tests on normality, overdispersion, and outliers of the full binary outcome model.

3.4 Discussion

The chapter aims to address the challenge in multilevel modelling when a unique valued student outcome is absent. In particular, this chapter aims to answer to the following two questions: (a) how to optimise reliability when fitting categorical outcome variables in a continuous outcome model; and (b) how to reliably use dichotomised continuous outcome variables in a binary outcome model?

In this chapter, I developed and compared three outcome models with different valued student outcomes under the same set of covariates. The three outcome models considered in this chapter are: (a) continuous outcome model where the outcome variable is the norm-adjusted scale scores following a normal distribution and the link function is identity; (b) categorical outcome model where the outcome variable is stanine score following approximately a normal distribution and the link function is also identity; and (c) binary outcome model where the outcome variable is binary following a Bernoulli distribution and the link function is logit.

Three full outcome models are examined in their model reliability and model accuracy. With respect to the research question of this chapter, it is found that: (a) all three full outcome models improve model goodness-of-fit in comparison to their respective empty and baseline models; (b) the full continuous outcome model using norm-adjusted scale scores produced the largest reduction in deviance proportional to those of its empty model; (c) the full binary outcome model has a singular fit; and (d) model estimations are consistent in direction and statistical significance between the full continuous and the full categorical models, while estimations from the full binary model are more conservative.

Based on these integrated results, from a statistical perspective, the full continuous outcome model is the preferred model as it best describes the given data and explains more variance of its empty model.

The full continuous and full categorical models although more efficient in explaining variance, require greater inference post-analysis around student learning. For example, an effect of six scale units shows an improvement in learning, but what this means requires interpretation based on the norm-referenced sample. For example, it is expected that students from Year 4 are expected to improve in term of scale units by six star units a year, therefore, extrapolating from this, we could conclude that an effect of six star units equals a year's growth in addition to expected growth. However, students from Year 6

are expected to progress eight star units in a school year. The effect of the intervention is measured uniformly across year levels, but the interpretation is not. It implies two things: either the model is measured uniformly with different interpretations by year level, or the model accounts for interactions in year levels across levels of predictors. The addition of cross-level interactions complicates the model specifications which are likely to become penalised by the AIC criteria.

The full binary outcome model has a singular fit, but it does not affect the estimated coefficients, rather it penalises the goodness-of-fit. However, the binary outcome model renders adequate goodness-of-fit indicating the reliability of its model estimation. Moreover, the estimated effects in the full binary model are more conservative than estimates rendered by the other two full models. It means that the presence of (even weakly) significant effects under the full binary outcome model can mean a greater practical significance.

The full binary model is also easier to interpret because the results will show either the criteria was met or not. For example, an odds ratio above 1 indicates improved learning outcomes in that students have higher odds of meeting the criteria. Moreover, the estimated effects in terms of odds ratios can be interpreted uniformly across year levels.

This finding has a practical implication for modelling: a uniformed model for a longitudinal or cross-sectional study where the absence of a unique outcome variable can be resolved using straightforward transformations on the outcome variable.

It provides evidence for how to transform outcome variables from different assessments for longitudinal comparison. Despite the difference in the degree of variance explained, each full model can detect intervention effects; meaning that they are all appropriate for their purpose. Therefore, studies involving learning outcomes at different year levels measured using different outcome variables can be integrated by transforming them into a single measure to compare longitudinally. In New Zealand, this has particular importance because the early measures at primary schools tend to be categorical (e.g., PM reading and Running Records), while the latter measures at primary school and early secondary are continuous (e.g., e-asTTle and PATs), while at the secondary level under NCEA, the outcomes are binary (e.g., achieved level 1 certificate) and categorical (e.g., endorsements in subjects). It means that we can develop longitudinal comparisons across multiple measures from primary to secondary through transformations into binary and or categorical variables. It overcomes existing limitations in the study where

dichotomisations of outcome variables are strongly discouraged (Altman & Royston, 2006).

Furthermore, it provides evidence that it is feasible to obtain reliable and accurate model fit when a categorical outcome is modelled linearly as how one would fit a continuous outcome. For example, stanines, widely used in standardised assessments, are categorical by nature, but also follows an asymptotic normal distribution. This finding suggests that in studies where stanine scores from different assessment tools are available, (a) it can serve as an appropriate valued student outcome in evaluating intervention success and (b) it can be treated as a continuous outcome which means complex models such as logistic or probit regressions can be avoided.

4 A Secondary School Case study

In the previous chapter, I addressed the challenge of the absence of a unique outcome measure in multilevel modelling through a case study of longitudinal school interventions in two clusters of New Zealand primary schools. Another challenge in developing multilevel models in intervention studies arises from the complexity of the underlying data structure.

The data for this project was part of a larger intervention which was conceptualised and designed by Wilson et al. (2017), with the statistical modelling component being developed by me for my PhD. The background of the study and the methods were published in Wilson et al. (2017) and are paraphrased here so that the reader understands the context of the case study. The statistical modelling described here extends the study by comparing models with varying specifications of random effects.

NCEA data as the outcome measure (and the only national qualification framework in New Zealand secondary school system) can complicate multilevel model specifications in evaluating intervention effectiveness. As discussed earlier in subsection 1.1.2, data structures are complex under the NCEA framework where (a) students can attempt various standards across and within subjects and qualifications (Level 1, 2, and 3) in a given academic year; (b) students can also repeat the same standard until they attain them across academic years and (in some cases) within the same academic year; (c) schools can design their own standard-based curriculums under the NCEA framework across subjects and qualifications; and (d) schools can also change their curriculum designs

across academic years and (in some cases) within the same academic year. This leads to situations where covariates which are usually considered as fixed effects render intraclass correlations such as subject and year level (Lai, et al., 2014; Lai, et al., 2018; Jesson, et al., 2018a; Jesson, et al., 2018b)

In this chapter, I address the challenge of complex data structures in multilevel modelling through a secondary school case study of a longitudinal intervention to facilitate attainments in a collection of achievement standards. Specifically, I examine models of four different specifications of fixed and random effects using the same student outcome.

I describe the study background and discuss the datasets and their structural complexity in Section 4.1. In Section 0, I provide four different model specifications and discuss their underlying assumptions. This is followed by comparisons of goodness-of-fit, variance partition coefficients, diagnostic tests of model assumptions, and model estimations in Section 4.3. Section 4.4 discusses the advantages and limitations of higher-level model specifications under these conditions of complex data structures.

4.1 Background to the study

This case study is part of a large-scale longitudinal intervention study that took place in secondary schools (n = 34) across two large regions in New Zealand. The intervention included the development of new forms of teaching in literacy, leadership enhancement and capability building in data use. Fourteen of the schools joined the project at the start of the study (*Group A*), and another 20 schools joined one year after (*Group B*). Schools predominantly drew students from communities with low to medium SES as measured in New Zealand by the *decile* index which is based on Census data for households with school-aged children in each school's catchment area. The *decile* index varies from 1 (*low* SES) to 10 (*high* SES) and is allocated by the New Zealand Ministry of Education to inform decisions about the funding schools receive. It is recalculated every six years.

One key research question is to build a value-added model to examine possible relationships between the levels and changes in literacy instruction, as measured through classroom observations, and pass rates in *SLAS* (see definition in subsection 1.1.2.2.2).

In the following subsections, in subsection 4.1.1, I describe participants involved in the wider intervention study. I then give a detailed account of measures collected in the study (subsection 4.1.2) before introducing the dataset used in this chapter. The complexity of the data structure is discussed in the following subsection, 4.1.3.

4.1.1 Participants

4.1.1.1 Schools

There were several limitations in the dataset. Due to practical constraints, such as school timetabling and researcher availability, not all schools were observed at both time points. The self-governing nature of the secondary school system in New Zealand, meant that not all schools offered all the standards deemed to have high literacy (*SLAS*) as part of their programmes. Thus, only those schools from which observation data were collected at both time points, and who had the *SLAS* outcome data are included. The final group of schools involved in these analyses included 7 *Group A* schools that participated in the literacy component at the end of the first year of interventions and 15 *Group B* schools that participated in the literacy component at the end of the second year of interventions.

Based on 2010 *decile* ratings, a total of eight schools drew children from decile 1-2 (*lowest SES*) areas, eight schools were located in decile 3-4 (*medium SES*) areas, and six schools were located in decile 5-8 (*higher SES*) area. Participating schools had relatively high proportions of students who identified as Māori or Pasifika (prioritised ethnicity) compared to national (see Table 4.1).

Table 4.1 Ethnicity Distribution of Participating Schools in Comparison with National Distribution in the Secondary School Case Study

Ethnicity	Case study	National
Māori	25.1%	18.3%
Pasifika	31.1%	9.0%
NZ European	27.1%	57.1%
Asian	12.0%	10.1%
Other	4.7%	6.2%

4.1.1.2 Teachers

A total of 212 Year 12 teachers participated in classroom observations voluntarily and 252 different lessons in classes were observed across two time points, (n = 90 English lessons, n = 88 mathematics lessons, and n = 74 science lessons). All teachers gave informed consent for the classroom observation.

4.1.2 Measures

4.1.2.1 Valued student outcome

A total of seven NCEA level 2 standards from three subjects, i.e., English, mathematics, and science were selected and named subject literacy achievement standards (*SLAS*) (see Table 4.2). National subject experts identified the *SLAS* as demanding high levels of literacy to achieve and were the principal measure of valued student outcome. Alignment of NCEA level 2 standards with The New Zealand Curriculum (2007) was drafted and finalised in 2011, and the re-aligned standards were in use in 2012. Table 4.2 shows the SLASs standard number before (*Expired*) and after (*Replacement*) the re-alignment. There are four ordered grades for NCEA achievement standards: Not Achieved (N), Achieved (A), Achieved with Merit (M), and Achieved with Excellence (E).

When the response variable is categorical and ordinal, as in this case study, several generalised logistic models can be applied, for example, ordinary polytomous model, cumulative odds model, continuation ratio model, and adjacent categories models (Agresti, 2013; Guo & Zhao, 2000; Jimenez & Salas-Velasco, 2000; Long, 1997). The trade-off between incorporating the natural ordering of grades and ease of interpretation of the model parameters was solved by simplifying the ordinal response variable to a dichotomous one. Therefore, the *SLAS* achievement results were collapsed into binary forms, i.e., not passed (N) or passed (either A, M, or E). The resulting binary achievement measure is the valued student outcome for this case study.

National pass rates were aggregated for each standard and each academic year across gender, ethnicity, and *decile* to serve as benchmarks for *SLAS* achievements.

Table 4.2 List of Subject Literacy Achievement Standards (SLAS)

Subject	Expired	Replacement
English	90377	91098
	90380	91100
Mathematics	90290	91261
	90292	91267
Science	90255	91171
	90308	91164
	90459	91157

4.1.2.2 Quality of literacy instruction

For the purpose of the in-class assessment of the quality of literacy teaching, an observation tool was developed to assess five core dimensions of this construct. It used an interval-based procedure with predefined categories as well as a running record which elaborated or exemplified the judgements made (Wilson, et al., 2017).

The dimensions were (Wilson, et al., 2017):

- Vocabulary: instructions which refer directly or by association to specific vocabulary;
- Structure: global structures of texts as well as structural/organisational and typographical features, e.g., headings, topic sentences, use of bolding;
- Audience/Purpose: reflections on what a text sets out to achieve with regard to the audience addressed;
- Language resources: function and impact of part of speech, grammar, rhetoric;
- Spelling and Punctuation.

Observers went into the classrooms and undertook observations in observing-coding cycles, intermittently focussing on the lesson for 3 minutes and coding observations for another 3 minutes. This resulted in multiple observational blocks, comprising three minutes of class time each, within each observed lesson. Between 7 and 10 three minutes blocks were typically collected in each class covering about 50% of time spent in classrooms. Each block was coded for the presence or absence of the literacy dimension judged by observing the teacher. The average of the five dimensions was calculated across observations within each English, mathematics and science department, resulting in a

subject index of the quality of literacy teaching (*LD-Coverage*). Across two time points, observers collected a total of 2283 blocks.

4.1.2.3 Case study data for model comparisons

The objective of this chapter is to study model robustness in terms of model reliability and accuracy under different model specifications of fixed and random effects. As a result, it is important to compare model specifications and behaviours under an identical dataset. In this subsection, I describe a subset of the case study data that is used in this chapter (see Table 4.3).

SLAS achievement data of all listed standards in Table 4.2 is the chosen valued student outcome. Covariates include student ethnicity (*E*), standard-specific national pass rates (*NPR*) at each academic year, subject of the standard (*S*), subject-specific index of *LD-Coverage* (*LDC*) before and after the intervention, school *decile* (*D*), number of years before (negative integer) and after (positive integer) the intervention (*T*), and number of years of intervention exposure (*I*).

Furthermore, I considered only Year 12 students from each school who participated in any *SLAS* standards over a period of five years. In a particular year, students can participate in none or all the *SLAS*s standards, and it is assumed that students do not repeat a school year in all participating schools.

The confined dataset reduces the data complexity where repeated measures of student across time are eliminated. However, covariates such *LD-Coverage* is a repeated measure across time at the school-subject level.

Table 4.3 Descriptions of Model Outcome and Covariates in the Secondary School Case Study

Variable	Type of Data	Description
Valued student outcome		
SLAS	Binary	Pass or not pass (an achievement standard)
Covariates		
Ethnicity (*E*)	Categorical	Māori (baseline), Pasifika, Others (includes NZ European and Asian)
National Pass Rate (*NPR*)	Continuous	Standard-specific pass rate by academic years, mean of 61.2%
Subject (*S*)	Categorical	English (baseline), Mathematics, Science
LD-Coverage (*LDC*)	Continuous	Subject index of the quality of literacy teaching, mean of 0.65 and standard deviation of 0.30 (number of blocks)
Decile (*D*)	Categorical	Lowest (Decile 1-2), Medium (Decile 3-4), Highest (Decile 5-8, baseline)
Time (*T*)	Continuous	Number of years before and after the intervention, ranges from -3 to 2
Intervention (*I*)	Continuous	Number of years of intervention exposure, ranges from 0 to 2

The outcome measure is in a binary form, and covariates include both categorical and continuous data. Transformations of data include: (a) dichotomisation of the outcome (as discussed earlier in subsection 4.1.2.1); (b) aggregations of New Zealand European, Asian, and others in ethnicity due to over-representation of the prioritised ethnicity groups (see Table 4.1); (c) categorisations of the school *deciles* (as discussed earlier in subsection 4.1.1.1).

The number of years before and after the intervention and number of years of intervention exposure are considered as continuous variables in the model in that linear cumulation of intervention effects are assumed.

4.1.3 Data structure

In this subsection, I discuss the complexity of the data structure presented in the case study and compare four possible structures by their underlying assumptions (see Table 4.4 where "~" means "nested within").

The data in this case study has two hierarchical structures where (a) repeated outcome measures across standards (and therefore subjects) are nested within students which are nested within schools and (b) subject-specific index of the literacy teaching quality are nested within schools. The two hierarchical structures are parallel in that students are not nested within subject because students can attempt standards across all three subjects.

This leads to the following three assumptions:

- A: Independence of repeated measures across standards and subjects
- B: Independence of students participating in the same standard
- C: Independence of students participating in the same subject

Table 4.4 Comparisons of Four Possible Structures by Assumptions in the Secondary School Case Study

Data Structure	Description	Assumption
2-level	Students ~ schools	A, B and C
3-level-D	Students ~ schools and Subject ~ schools	A and B
3-level-R	Repeated measures (across standards) ~ students ~ schools	C
4-level	Repeated measures (across standards) ~ students ~ schools and Subject ~ schools	-

Assumption A is that a student's chance of passing each standard is independent of each other. This assumption is unlikely to be met where the odds of passing a *SLAS* standard is likely to increase in the case where attainments in other *SLAS* standards are achieved, especially for standards of the same subjects.

Assumption B is that there is no underlying correlation among students participating in the same standards whereas assumption C is that there is no underlying correlation among students participating in standards from the same subject. Assumption B contains C in that violation of B implies violations of C, but the reverse may not hold.

In this chapter, I compare multilevel models using model specifications under the four assumptions described in Table 4.4: *2-level, 3-level-D, 3-level-R,* and *4-level*. The hierarchical *2-level* structure assumes all three assumptions, while none is assumed in the *4-level* structure where the two nested structures are both defined. The *3-level-R* structure extends the hierarchy of the *2-level* structure by adding repeated measures across standards, as a result, reduces model assumptions to only independence of within-subject participants. The *3-level-D* structure, on the other hand, extends the *2-level* model in an alternative way by including the parallel structure presented in the covariates.

Table 4.5 Summary of Model Specifications of Fixed and Random Effects by Data Structures in the Secondary School Case Study

Covariate	2-level	3-level-D	3-level-R	4-level	Temporal
Random Effects					-
Level 1	Achievement	Achievement	Achievement	Achievement	-
Level 2	School	Subject	Student	Student	-
Level 3	-	School	School	Subject	-
Level 4		-	-	School	-
Fixed Effects					
National Pass Rate (NPR)	Level 1	Level 1	Level 1	Level 1	Academic Year
Subject (S)	Level 1	Level 2	Level 1	Level 3	Constant
LD Coverage (LDC)	Level 1	Level 1	Level 1	Level 1	Intervention Year
Ethnicity (E)	Level 1	Level 1	Level 2	Level 2	Constant
Time (T)	Level 1	Level 1	Level 2	Level 2	Academic Year
Intervention (I)	Level 1	Level 1	Level 2	Level 2	Intervention Year
Decile (D)	Level 2	Level 3	Level 3	Level 4	Constant

4.2 The models

Following the previous subsection, I describe the model specifications in terms of fixed and random effects in this subsection, first in English then in mathematics.

There are four model specifications considered in this chapter: the *2-level* model, *3-level-D* model, *3-level-R* model, and *4-level* model. The hierarchical *2-level* model takes accounts of correlations among participated achievement standards within schools. The *4-level* model improves the *2-level* model by modelling (a) repeated measures across standards within students i.e., the student-specific random effect, as well as (b) similarities potentially rendered by the cluster of students taking the same standards, i.e., subject-specific random effect. The *3-level-D* model and *3-level-R* model are two partitions of the *4-level* model where the former examines only the subject-specific effect and the latter fits only the student-specific effect.

A summary of each model specification of fixed and random effects is shown in Table 4.5. It is apparent that the *2-level* model and the *4-level* model serve as the lower and upper limits of the model fits. As a result, trade-offs between fitting only the subject-specific effect and only the student-specific effect can be made by comparing model reliability and accuracy between the two models with three levels.

The `school` variable is considered as a random effect across all four specifications, specifying the multilevel structure of school to account for correlations in the *SLAS* achievements due to the nesting of students within schools. `Student` (as in unique National Student Numbers (`NSN`) in the data) is fitted as an additional random effect in the *3-level-R* model and the *4-level* model. The interaction term of `School` and `Subject` were also fitted as a random effect accounting for the subject-specific effect is included in the *3-level-D* model and the *4-level* model.

The goal of the original research question (of the intervention study) is to investigate the effectiveness of the literacy program at the school-subject level. The quality of literacy instruction measure, *LD-Coverage* is the key covariate of interest. There is an underlying temporal structure in relation to intervention years within the *LD-Coverage* measure in that two repeated measures were collected for each subject within each school: one before, and the other after, the literacy intervention.

Another covariate that has an underlying temporal structure is the *national pass rate*, which is used to control for national trends in student pass rates. Another way of

interpreting it is to consider it a measure of the degree of difficulties where high national pass rates imply a low level of difficulty. The standard-specific national pass rates for each academic year (`NPR`) were introduced as a covariate at the lowest level across all four model specifications. Although each model estimates individual-specific effects of national pass rate, their underlying assumptions of independence vary (see Table 4.4), as a result, estimations may be biased when underlying assumptions are not satisfied.

Accounting for those temporal structures of the repeated data in the covariates, a fixed effect of time (`Time`) measured in years of participation was introduced and centred around the baseline year of the intervention. A second temporal covariate is introduced (`Intervention`) to allow for a discontinuous slope of time-related to the intervention (0 before the intervention; the number of years of intervention exposure after the intervention, see Singer & Willett, 2003, p. 195-198). Both temporal measures are fitted as fixed effects.

Moreover, to test for the hypothesised impact of the intervention on the quality of literacy instruction in classrooms, and, in turn, on student pass rates, the interaction effect `Intervention` × `LD-Coverage` was introduced to allow for a change in slope in relation to changes in *LD-Coverage*.

I hypothesised that higher *decile* schools would have higher pass rates, so *decile* was introduced to each model as a school-level fixed effect in the form of a categorical variable `Decile`.

To account for variations at student-level, ethnicity (`Ethnicity`) was included in each model as a student characteristic.

To allow for subject-specific intercepts, the *subject* was introduced as another fixed effect (`Subject`) in each model specifications.

In the next four subsections, I describe each model in mathematics following notations and structures presented in Section 2.1. Although a complete understanding of the following subsection is not required, I present here model specifications in mathematics for its completeness and robustness.

4.2.1 Logit link function

Because student achievement data are in binary format (1 = passed, 0 = not passed), the logit link function is applied in all four models. Let Y_{ijklt} be the binary variable with

Bernoulli distribution $\Pr(Y_{ijklt} = 1) = p$, where index *i* represents the i^{th} student, index *j* represents the j^{th} standard, index *k* represents the k^{th} subject, index *l* represents the l^{th} school, index *kl* represents the kl^{th} school subject, and index *t* represents time measured in participation year. The logit link function is defined as:

$$\text{logit}\,\mathrm{E}(Y_{ijklt}) = \text{logit}\,\Pr(Y_{ijklt} = 1) = \log\left(\frac{p}{1-p}\right) = \eta_{ijklt} \qquad (4.1)$$

where p is unknown parameter and η_{ijklt} is the linear predictor of its log-odds.

For ease of reading, in the following subsections, I define model specifications on η_{ijklt} instead of Y_{ijklt} where their linkage is defined in Equation 4.1. Notations of covariates used in the following subsections are given in Table 4.3. The assumptions of each model are described in Section 2.2 and the variance-covariance structures follow the multivariate assumptions defined in subsection 2.1.2.

As discussed in Section 2.2, residual variances from models using logit link function with Bernoulli distribution are normalised so that level-1 residual variance is constant at approximately 3.29. As a result, direct comparisons of random effects should be avoided between different models. Model specifications of fixed and random effects are summarised in Table 4.5.

4.2.2 The *2-level* model

The *2-level* model is defined as:

$$\eta_{ijklt} = \beta_{0l} + \beta_1 NPR_{jt} + \beta_2 S_k + \beta_3 LDC_{klt} + \beta_4 E_i + \beta_5 T_i + \beta_6 I_i \qquad (4.2)$$

$$\beta_{0l} = \gamma_{00} + \gamma_{01} D_l + \nu_{0l} \qquad (4.3)$$

where:

- β_{0l} = intercept for school l,
- β_1 ... β_6 = regression coefficient associated with each level-1 covariate,
- D_l := school decile for school l,
- γ_{00} = overall mean intercept adjusted for school deciles,
- γ_{01} = regression coefficient associated with school deciles with respect to level-1 intercept,
- ν_{0l} = random effects of school l adjusted for school deciles on the intercept.

4.2.3 The *3-level-D* model

The *3-level-D* model is defined as:

$$\eta_{ijklt} = \beta_{0kl} + \beta_1 NPR_{jt} + \beta_2 LDC_{klt} + \beta_3 E_i + \beta_4 T_i + \beta_5 I_i \quad (4.4)$$

$$\beta_{0kl} = \gamma_{00l} + \gamma_{01} S_k + \nu_{0kl} \quad (4.5)$$

$$\gamma_{00l} = \gamma_{000} + \gamma_{001} D_l + \nu_{00l} \quad (4.6)$$

where:

- β_{0kl} = intercept for subject k in school l,
- $\beta_1 \ldots \beta_6$ = regression coefficient associated with each level-1 covariate,
- $S_k :=$ subject for each subject k in school l,
- γ_{00l} = mean intercept for school l adjusted for the subject,
- γ_{01} = regression coefficient associated with the subject with respect to level-1 intercept,
- ν_{0kl} = random effects of subject k in school l adjusted for the subject on the intercept,
- $D_l :=$ school decile for school l,
- γ_{000} = overall mean intercept adjusted for school deciles,
- γ_{001} = regression coefficient associated with school deciles with respect to level-2 intercept, and
- ν_{00l} = random effects of school l adjusted for school deciles on the intercept.

4.2.4 The *3-level-R* model

The *3-level-R* model is defined as:

$$\eta_{ijklt} = \beta_{0il} + \beta_1 NPR_{jt} + \beta_2 LDC_{klt} + \beta_3 S_k \quad (4.7)$$

$$\beta_{0il} = \gamma_{00l} + \gamma_{01} E_i + \gamma_{02} T_i + \gamma_{03} I_i + \nu_{0il} \quad (4.8)$$

$$\gamma_{00l} = \gamma_{000} + \gamma_{001} D_l + \nu_{00l} \quad (4.9)$$

where:

- β_{0il} = intercept for student i in school l,
- $\beta_1 \ldots \beta_3$ = regression coefficient associated with each level-1 covariate,
- γ_{00l} = mean intercept for school l adjusted for level-2 covariates,
- $\gamma_{01} \ldots \gamma_{03}$ = regression coefficient associated with level-2 covariates with respect to level-1 intercept,

- v_{0il} = random effects of student i in school l adjusted for level-2 covariates on the intercept,
- D_l := school decile for school l,
- γ_{000} = overall mean intercept adjusted for school deciles,
- γ_{001} = regression coefficient associated with school deciles with respect to level-2 intercept, and
- v_{00l} = random effects of school l adjusted for school deciles on the intercept.

4.2.5 The *4-level* model

The *4-level* model is defined as:

$$\eta_{ijklt} = \beta_{0\mathrm{ikl}} + \beta_1 NPR_{jt} + \beta_2 LDC_{klt} \quad (4.10)$$

$$\beta_{0ikl} = \gamma_{00\mathrm{kl}} + \gamma_{001} E_i + \gamma_{002} T_i + \gamma_{003} I_i + v_{0\mathrm{ikl}} \quad (4.11)$$

$$\gamma_{00\mathrm{kl}} = \gamma_{000\mathrm{l}} + \gamma_{001} S_k + v_{00kl} \quad (4.12)$$

$$\gamma_{000\mathrm{l}} = \gamma_{0000} + \gamma_{0001} D_l + v_{000l} \quad (4.13)$$

where:

- β_{0ikl} = intercept for student i in the subject k in school l,
- β_1, β_2 = regression coefficient associated with each level-1 covariate,
- γ_{00kl} = mean intercept for subject k in school l adjusted for level-2 covariates,
- $\gamma_{001} \ldots \gamma_{003}$ = regression coefficient associated with level-2 covariates with respect to level-1 intercept,
- v_{0ikl} = random effects of student i in subject k in school l adjusted for level-2 covariates on the intercept,
- S_k := subject for each subject k in school l,
- γ_{000l} = mean intercept for school l adjusted for the subject,
- γ_{001} = regression coefficient associated with the subject with respect to level-2 intercept,
- v_{00kl} = random effects of subject k in school l adjusted for subject on the intercept,
- D_l := school decile for school l,
- γ_{0000} = overall mean intercept adjusted for school deciles,
- γ_{0001} = regression coefficient associated with school deciles with respect to level-3 intercept, and

- ν_{000l} = random effects of school l adjusted for school deciles on the intercept.

Each model specification (i.e., fixed effect coefficients βs and γs and random effects νs) is then estimated based on the observed data using `R` (Version 3.5.4, R Core Team, 2018) with the `lme4` package (Bates, Maechler, Bolker & Walker, 2015a, 2015b). Diagnostic tests of model fit are performed by the `DHARMa` package (Hartig, 2019) is used to check for model assumptions based on simulation. Model goodness-of-fit are assessed using AIC and deviance as described in Section 2.3 and likelihood ratio tests are applied for model comparisons. I present in the next section, the results of each model described in this section.

4.3 Results

In this section, I compare the empty models across the four model specifications i.e. the *2-level*, *3-level-D*, *3-level-R*, and *4-level* models in subsection 4.3.1. In subsection 4.3.2, I compare the full models across four model specifications followed by diagnostics tests of models assumptions of each full model in subsection 4.3.3. For a complete set of model estimations, please see Appendix J to Q.

4.3.1 Empty models

Following the analysis strategy set in Section 2.4, I first compare the empty models across four model specifications where in addition to the individual-specific effect (which is normalised at 3.29 in all empty models): (a) the *2-level* model including the school-specific effect; (b) the *3-level-D* model including both subject- and school-specific effects; (c) the *3-level-R* model including both student- and school-specific effects; and (d) the *4-level* model including student-, subject-, and school-specific random effects.

Comparisons between nested empty models such as (the empty 2-level model and the empty 3-level-D model or the empty 3-level-R model and the empty 4-level model) can be made directly to determine the significance of the additional random effect (included in the higher level).

Table 4.6 summaries model estimations of the empty model under each model specification, as well as statistics of the likelihood ratio tests across models. Likelihood ratio tests were performed on two parallel stepwise-nested models: (1) *2-level ~ 3-level-D ~ 4-level* and (2) *2-level ~ 3-level-R ~ 4-level*. However, the difference in deviance between the two 3-level models cannot be statistically tested using the likelihood ratio test because they are not sequentially nested.

Table 4.6 shows that the *2-level* model is significantly improved by its sequential 3-level models where both chi-squared statistics are significant, and AICs are lower. Likewise, the *4-level* model (that includes both the student- and subject-specific effects) improves the two 3-level models at 5% significance, which is indicated by highly significant chi-squared statistics and much lower AICs.

The addition of the subject-specific effect alone increased the estimated fixed effect, i.e., intercept (from 0.29 to 0.31), while the inclusion of the student-specific effect alone reduced the estimated fixed effect (from 0.29 to 0.26). The *4-level* model renders the lowest estimation of the fixed effect at 0.23.

As discussed earlier in subsection 2.1.1.3, in multilevel models using the logit link function (and many other GLMMs), the value of the regression coefficients and higher-level variances are rescaled to keep the lowest-level residual variance constant at $\frac{\pi^2}{3}$ (Snijders and Bosker, 1999). As a result, it is infeasible to compare regression coefficients and changes in variance components across the four empty models (Snijders and Bosker, 1999).

Similar to the previous chapter (see in subsection 3.3.1), I discuss here the variance partition coefficient as an intraclass correlation coefficient where the cut-off point at 4% for reliable estimations and 7% for excellent estimations of group-level effects are adapted (Fleiss, 1986).

It is estimated that around 8.1% of the total variance is due to the variance between schools under the *2-level* model (see Table 4.7). It can be read as the average correlation between two randomly selected standard achievements from the same school is expected to be 0.81. Estimations of the school-level intraclass correlation coefficients are approximately 8% across all four model specifications. Those results suggest that estimations of school-specific effects are excellent across all four models.

Table 4.6 Summary of the Empty Models in the Secondary School Case Study

Variance of	2-level	3-level-D	3-level-R	4-level
Fixed Effects				
Intercept (s.e.)	0.29 (0.10)	0.31 (0.10)	0.26 (0.12)	0.23 (0.13)
Random Effects				
Repeated Measure	3.29	3.29	3.29	3.29
Student	-	-	1.39	1.55
Subject	-	0.12	-	0.22
School	0.29	0.29	0.44	0.45
Likelihood Ratio Test				
AIC	54029	53647	51998	51468
Deviance	54019	53641	51992	51460
df	2	3	3	4
χ^2	-	377.1	2026.2	2181.1[1] | 532.1[2]
P-value	-	< .001	< .001	<.001[1] | < .001[2]

1: test statistics against *3-level-D* model and 2: test statistics against the *3-level-R* model.

Table 4.7 Variance Partition Coefficient of Each Empty Models in the Secondary School Case Study

Level	2-level	3-level-D	3-level-R	4-level
Repeated Measure	91.9%	88.9%	64.3%	59.7%
Student	-	-	27.1%	28.1%
Subject	-	3.3%	-	4.0%
School	8.1%	7.8%	8.6%	8.2%

The large variation remained at the lowest level (i.e., individual achievement in standards) in the *2-level* model (91.9%) is partially due to the violation of the independence assumption where two randomly selected individual achievements in standards from the same school can be undertaken by the same student (i.e., assumption A in subsection 4.1.3).

In the *3-level-R* models with the student-specific effect, the intra-student correlation is estimated to be at 27%, which suggests that student-specific random effect is highly relevant to the model estimations of the *3-level-R* models.

On the other hand, when specifying an empty 3-level model with the subject-specific effect instead, the variance partition coefficient of the subject-specific effect is under the benchmark for reliable estimations (3.3%, which is below 4%).

However, in the empty 4-level model where both the student- and subject-specific effects are fitted, estimations of variance partition coefficients are all at/above the 4% benchmark indicating reliable estimations at each level of the model specification.

The empty models are not yet value-added models since none of the fixed effects is modelled (apart from the intercepts) (Goldstein, 1997). However, it renders the following patterns: (a) model fit improves when each additional random effect is fitted; (b) student-specific effect is more relevant to model estimations compared with subject-specific effects.

In the next subsection, I present and compare results of models in full specifications of the fixed and random effects (defined in Section 4.2).

4.3.2 Full models

The objective with respect to the original case study is to examine value-added effects by the improved quality of literacy instruction as well as the overall intervention, therefore the interaction of *LD-Coverage* and number of years in intervention is fitted in each full model where covariates such as national pass rate, subject, ethnicity, decile, and time are also controlled to reduce the misspecification bias so that the underlying intervention effects are estimated with reliability and confidence (Ferrão & Goldstein, 2009; Goldstein, et al., 2007; Foley & Goldstein, 2012).

The dataset used in this chapter is a small subset of a much larger intervention study where its underlying data structure is much more complex than the one used in here. The

case study dataset includes repeated measures of both valued-student outcomes and intervention-specific covariates (such as *LD-Coverage*) collected at various time points across schools.

It is not within the scope of this thesis to discuss the appropriateness or impact of the inclusion and/or exclusion of certain controlling variables. Instead, I focus in this thesis on a given set of covariates to demonstrate the trade-off between fixed and random effects of variables that define the data structure. For example, the covariate *subject* in this case study defines the data structure as it can represent the department from a school and therefore can be fitted as either a fixed or random effect, or both as in the case of the *3-level-D* and *4-level* models.

In the following subsection I compare model fits and estimations of the four final model specifications including (a) the *2-level* model with school-specific random effects; (b) the *3-level-D* model with both subject- and school-specific random effects; (c) the *3-level-R* model with both student- and school-specific random effects; and (d) the *4-level* model with student-, subject-, and school-specific random effects.

Table 4.9 presents the estimated fixed effects of each full model specification. When comparing the estimates in Table 4.9, the direction of the relationships is consistent for all fitted covariates across all four models. The statistical significance of the relationships is also maintained across four models except for decile and time. Schools in medium SES groups are estimated to be marginally different from schools in the highest SES (at 10% significance) in all models but the model with two levels (p-value = 0.11). Linear relationships between time and the linear predictor are estimated to be significant at 5% in all models except for the *3-level-R* model (at 10% significance).

Comparisons of change in magnitude are not feasible as estimated coefficients between different multilevel models with the logit link function are rescaled.

Comparisons between odds ratios (as effect sizes) for covariates across models should be made with cautions as they are conditional or group-specific measures of associations or intra-group measures of associations, not the population effect (Austin & Merlo, 2017). Take the overall *LD-Coverage* for example, in the *2-level* model, the odds ratio of 0.88 ($e^{-0.13}$) can be interpreted as when comparing two randomly selected performances in standards whose exposure to *LD-Courage* differs by 1 (classroom observation interval), but who share identical values on all other covariates and who also are within the same school, then the odds of passing are 0.88 times lower for those who have 1 unit of *LD-*

Coverage lower on average. That is the interpretation is conditional on both the other fixed effects as well as the random effects. In the *2-level* models, the odds ratio of the estimated *LD-Coverage* is conditional on all the covariates at the achievement level and the school-specific random effect. On the other hand, in the *4-level* model, the odds ratio of *LD-Coverage* (0.78) is conditioned on the covariates at four different levels adjusting for student-, subject-, and school-specific random effects (Austin & Merlo, 2017).

Larsen and Merlo (2005) described an approximation method of the population average coefficient for group-level covariates based on the conditional regression coefficient. Moreover, there is a collection of summary measures of the effects of group-level covariates available including interval odds ratio, the proportion of opposed odds ratio, median odds ratio, the proportion of variance explained by multilevel logistic regression model (Austin & Merlo, 2017). Nuanced comparisons of these summary measures are not discussed in this thesis as comparisons in direction and statistical significance are enough to determine the consistency and reliability of model estimations.

Although comparisons between estimated coefficients are infeasible, model comparisons can be made in terms of their goodness-of-fit through likelihood ratio tests. As for the empty models, likelihood ratio tests were performed on two parallel stepwise nested full models: (1) *2-level ~ 3-level-D ~ 4-level* and (2) *2-level ~ 3-level-R ~ 4-level*.
As shown in Table 4.9, the *2-level* model is significantly improved by its sequential 3-level models where both chi-squared statistics are significant, and AICs are lower. Likewise, the *4-level* model (that includes both the student- and the subject-specific effects) improves the two 3-level models significantly at 5%, which is indicated by highly significant chi-squared statistics and much lower AICs.

One final step to follow is checking model assumptions on linearity, normality, homoscedasticity, independence, and overdispersion. Model assumptions are examined in the following subsection both visually and in some cases statistically.

The prediction accuracy of each model is also computed based on a sample of testing data (based on another subset of the original data). It shows in Table 4.8 that overall accuracy of model predictions is around 82% for the *3-level-R* model and the *4-level* model, while for the *2-level* model and the *3-level-D* model, they are below 70%.

Moreover, the error rates of the false pass are extremely high in both of the *2-level* (72%) and *3-level-D* (70%) models where students are over-estimated in their odds of passing.

Error rates of false fail are controlled under 8% for both the *3-level-R* model (7.7%) and *the 4-level model* (7.6%).

Table 4.8 Model Accuracy Based on a Testing Sample in the Secondary School Case Study

Model	2-level	3-level-D	3-level-R	4-level
Accuracy	65.3%	66.3%	81.9%	82.3%
False Pass	72.1%	70.2%	35.2%	34.3%
False Fail	12.1%	11.7%	7.7%	7.6%

Table 4.9 Model Estimations of Fixed Effects of the Full Models in the Secondary School Case Study

Covariate	2-level	3-level-D	3-level-R	4-level
Fixed Effects	Coef. (s.e.)			
Intercept	-1.28 (0.21)	-1.18 (0.23)	-1.85 (0.25)	-1.70 (0.28)
National Pass Rate (NPR)	3.43 (0.31)	3.39 (0.31)	4.65 (0.36)	4.55 (0.37)
Subject: Mathmatics	-0.47 (0.05)	-0.51 (0.09)	-0.67 (0.05)	-0.71 (0.11)
Subject: Science	-0.59 (0.04)	-0.63 (0.09)	-0.96 (0.05)	-1.00 (0.11)
LD Coverage (LDC)	-0.13 (0.06)	-0.15 (0.07)	-0.22 (0.07)	-0.25 (0.09)
Ethnicity : Other	0.57 (0.03)	0.59 (0.03)	0.78 (0.05)	0.78 (0.05)
Ethnicity: Pasifika	-0.12 (0.03)	-0.12 (0.03)	-0.17 (0.06)	-0.17 (0.06)
Time (T)	0.05 (0.02)	0.06 (0.02)	0.04 (0.02)	0.06 (0.02)
Intervention (I)	-0.26 (0.04)	-0.31 (0.05)	-0.35 (0.06)	-0.42 (0.07)
Decile: Lowest	-0.69 (0.14)	-0.73 (0.14)	-0.88 (0.16)	-0.95 (0.17)
Decile: Medium	-0.21 (0.13)	-0.23 (0.14)	-0.29 (0.15)	-0.31 (0.17)
LD Coverage × Intervention	0.25 (0.07)	0.32 (0.07)	0.37 (0.08)	0.47 (0.09)
Likelihood Ratio Test				
AIC	53939	52624	50929	50751
Deviance	52913	52596	50901	50591
df	13	14	14	15
χ^2	-	317.2	2012.2	2004.6^1 | 309.6^2
P-value	-	< .001	< .001	$< .001^1$ | $<.001^2$

1: test statistics against *3-level-D* model and 2: test statistics against the *3-level-R* model.

4.3.3 Assumption checking

In this subsection, I present diagnostics of each full model through a combination of visualisation techniques and hypothesis testings. Comparisons are made using the `DHARMa` packaged in `R` based on simulations of residual errors from each model fit.

Scatter plots of the simulated standardised residual errors and fitted values for each full model is shown in Figure 8 to Figure 11. Linearity assumptions are met across all full models where the non-linear pattern is not found within each quartile of the residual errors against the predicted lines. Scatters across four figures show randomness indicating satisfactory of independence assumptions. Q-Q plots of residuals show strong and significant deviations from the standardised normal distribution in the full *3-level-R* and *4-level* models indicating violations of normality when student-specific effects are modelled.

Histograms of the standard deviations of the residuals are presented (at the centre) in Figure 12 to Figure 15 to inspect patterns of overdispersion visually. It shows that overdispersion presents in the *3-level-R* model (see Figure 14). The presence of overdispersion in the *3-level-R* model suggests that controlling for student-specific effect, there are slices of observed data where the probability of passing is closer to zero or one. It suggests that there is potentially a collection of high-performing students who attempted and passed all available standards or a collection of low-performing students who attempted but failed all available standards. The disappearance of the overdispersion in the subsequent *4-level* model using both student- and subject-specific effects suggests that the slices of observed data causing overdispersion are scattered randomly across subjects.

Histograms of simulated residuals are also shown in Figure 12 to Figure 15 (on the far right panels) where outliers are highlighted in red and are also tested based on the exact binary tests. In both *3-level-R* model and *4-level* model, influential outliers are found ($p < 0.1$). It is consistent with the findings on overdispersion in that (a) there is a cluster of high- or low- performing students (b) there may also be a cluster of high- or low-performing departments (although their variance is not significantly larger than the expected).

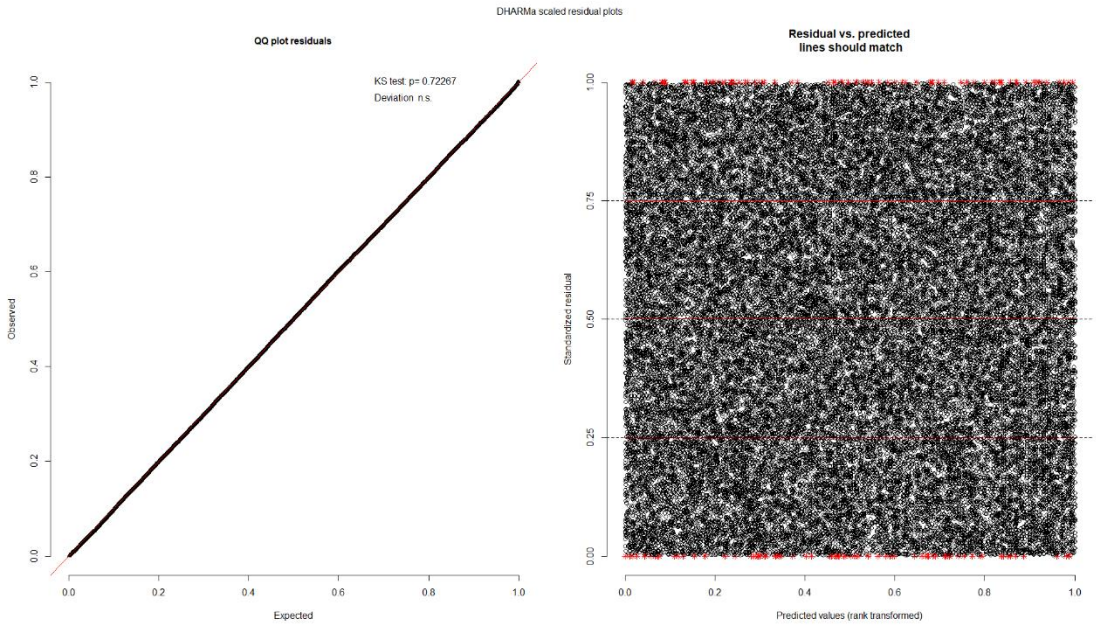

Figure 8 Residual versus fitted values and the residual normal Q-Q plot for the full 2-level model.

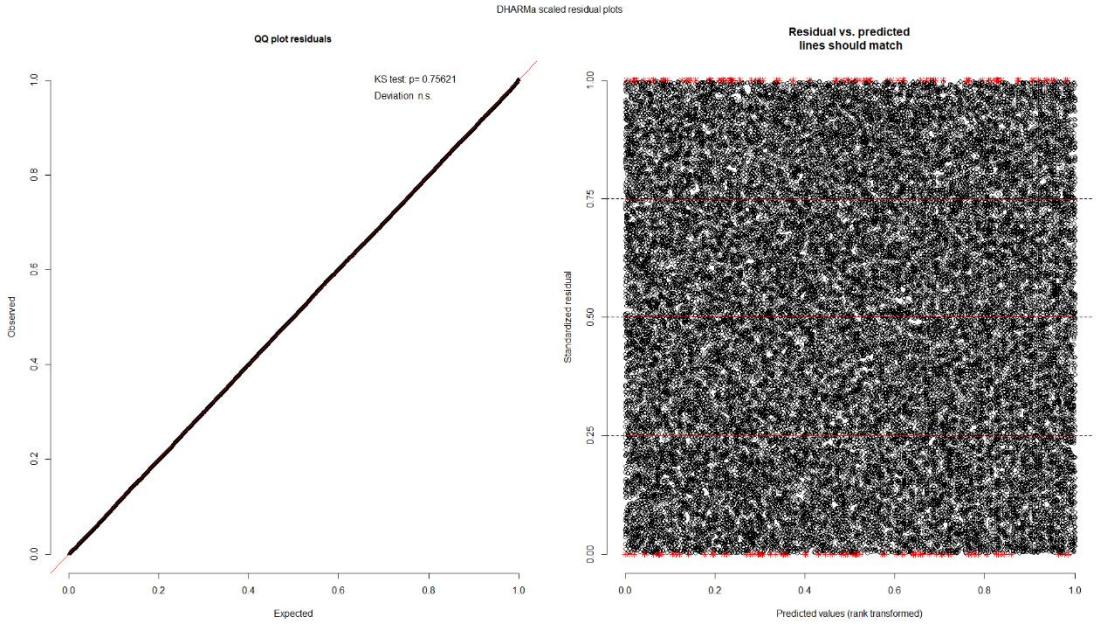

Figure 9 Residual versus fitted values and the residual normal Q-Q plot for the full 3-level-D model.

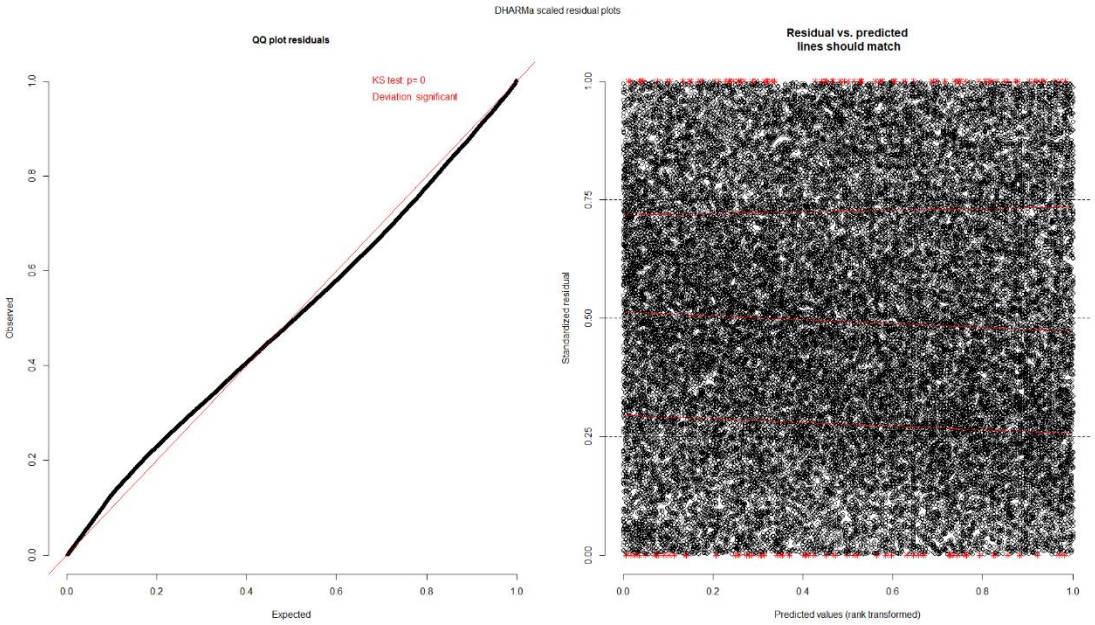

Figure 10 Residual versus fitted values and the residual normal Q-Q plot for the full 3-level-R model.

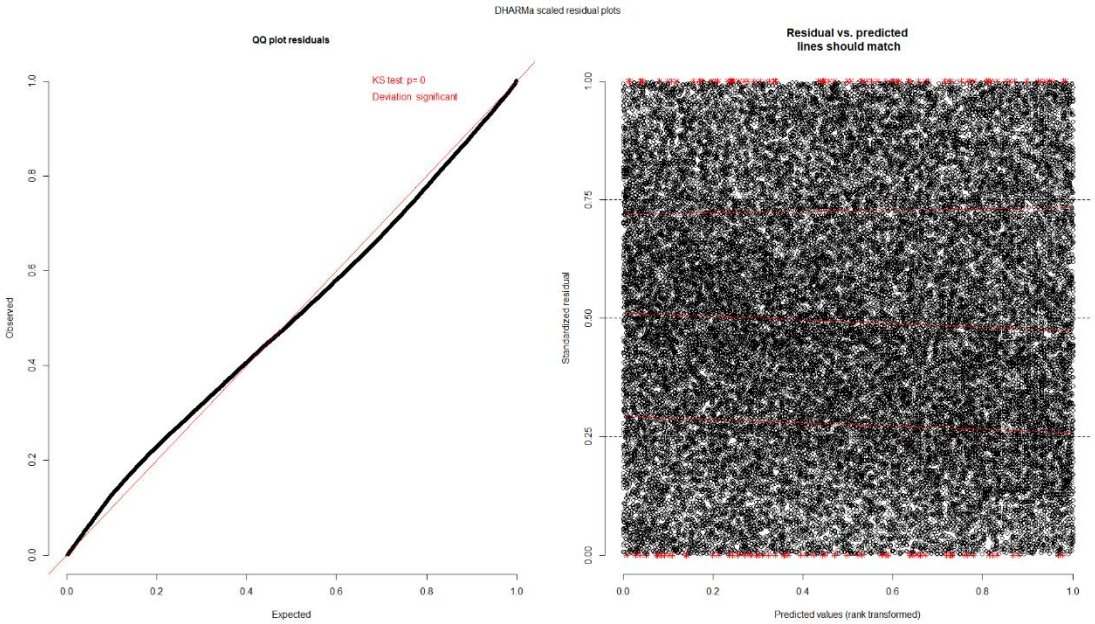

Figure 11 Residual versus fitted values and the residual normal Q-Q plot for the full 4-level model.

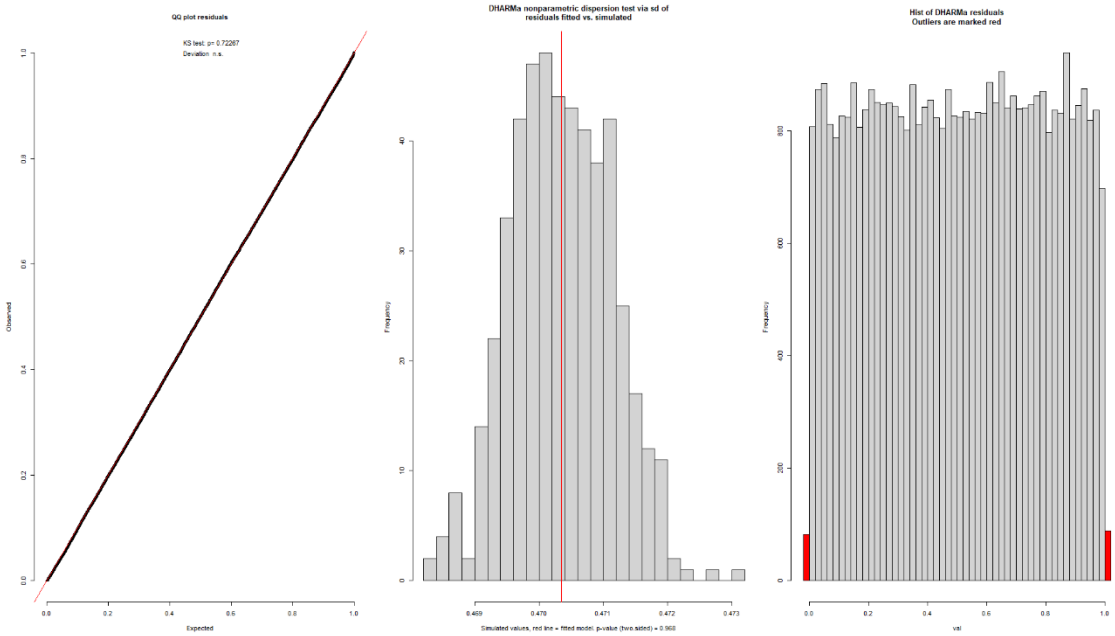

Figure 12 Tests on normality, overdispersion, and outliers for the 2-level model.

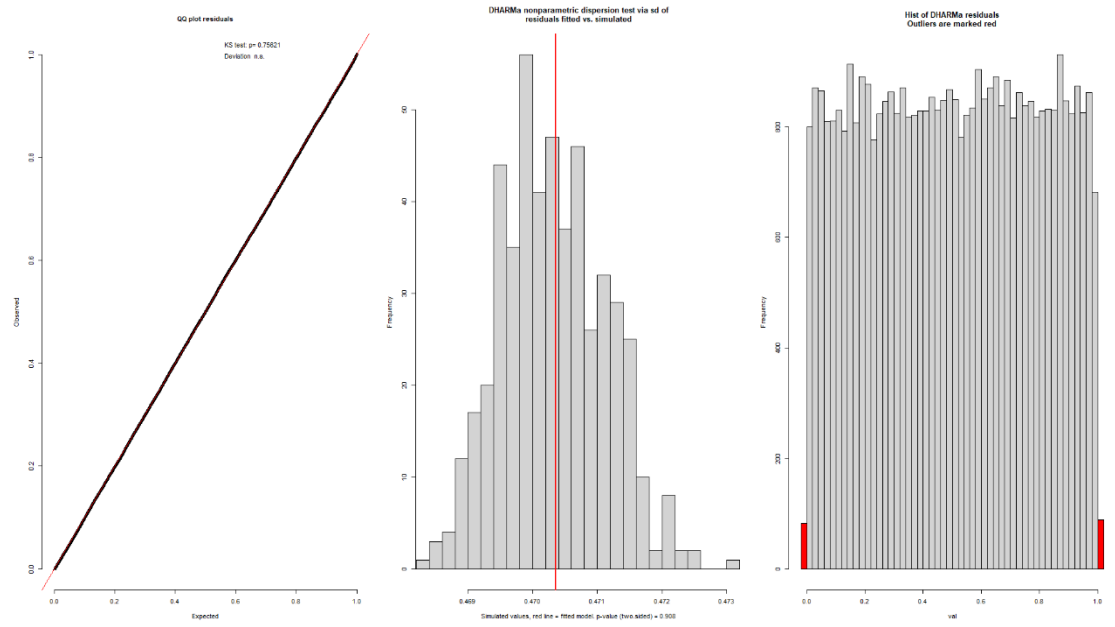

Figure 13 Tests on normality, overdispersion, and outliers for the 3-level-D model.

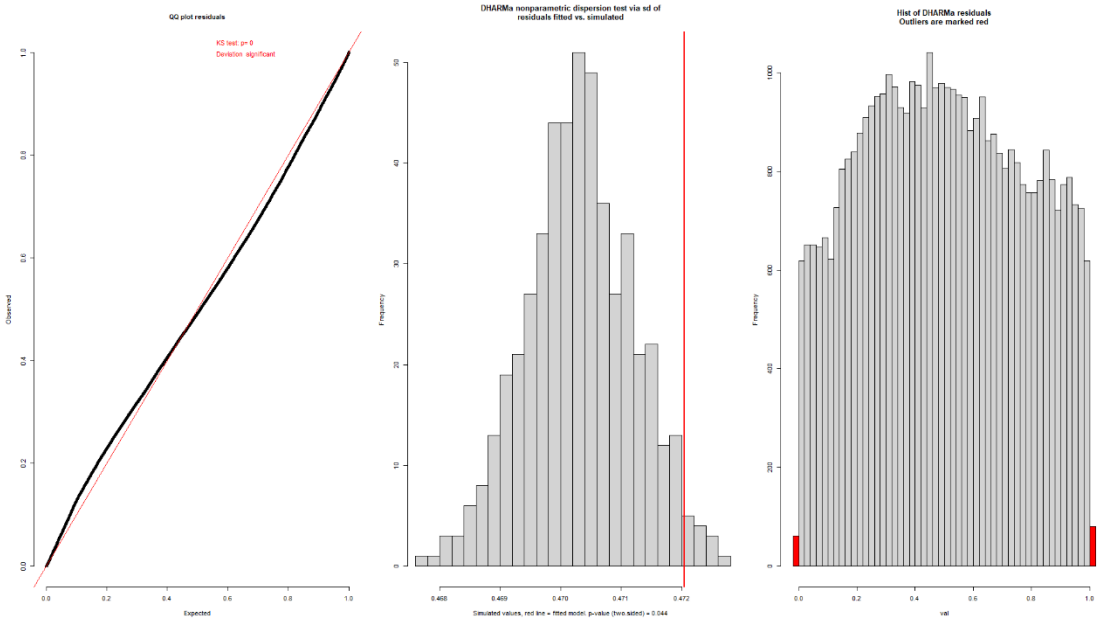

Figure 14 Tests on normality, overdispersion, and outliers for the 3-level-R model.

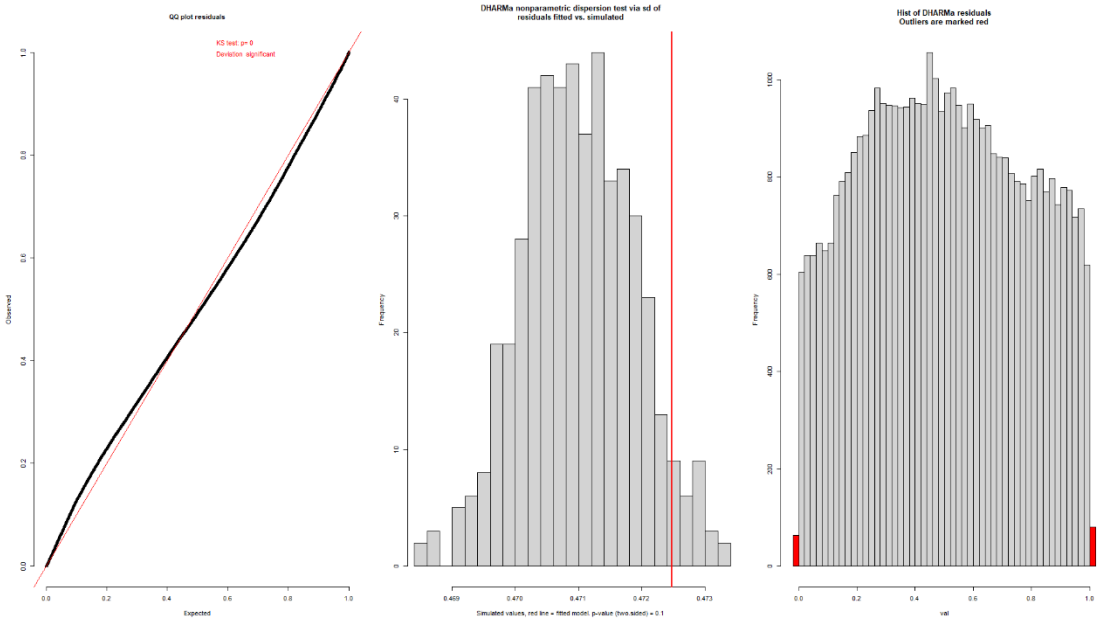

Figure 15 Tests on normality, overdispersion, and outliers for the 4-level model.

4.4 Discussion

This chapter aimed to address the challenge of the complex data structure in multilevel models with applications to binary outcome variables. Explicitly, it set out to answer the question of how many underlying random effects a multilevel model with the categorical (or binary) outcome should be fitted in order to reliably and accurately determine individual- and group-specific effects.

In this chapter, I examined the reliability and accuracy of various approaches to judging intervention outcomes in a complex standards-based environment. The usual approach to judging intervention effects in high school examination outcomes is to use fixed effects of such variables as year level and subject. However, NCEA creates complexities in that subjects vary across schools, years, year levels and students. Also, many school interventions are subject-related such as literacy instruction (Lai, et al., 2018; Lai, et al., 2014; Jesson, et al., 2018a; Jesson, et al., 2018b), which renders subject-specific intervention measures.

Four model specifications with varying degrees of random effects (at student-, subject-, and school-level) were compared in terms of model reliability (such as goodness-of-fit) and model accuracy (such as model assumptions) based on an intervention of improving literacy instruction quality is designed to improve pass rates in achieving *SLAS* standards in NCEA.

With respect to the research question of this chapter, it is found that: (a) student-specific random effects are a relevant source of variation in *SLAS* achievements; (b) subject-specific random effects can be an additional source of variation when student-specific random effects are also estimated; (c) model goodness-of-fit improves when an additional random effect is included; (d) model estimations are consistent in relation to direction and statistical significance across models; (e) fitting the student-specific random effects improves model accuracy more than fitting the subject-specific random effects.

As shown by the AIC criteria and the likelihood ratio tests, models with the student-specific random effect improve the reliability of the model estimations. Further inclusion of the subject-specific random effect improves model reliability further. From this perspective, the model with both student- and subject-specific random effects (as well as the usual school-specific effect) provides a more robust way of determining what worked (or did not work). The reasons for the greater robustness are to do with how we deal with

the assessment environment created by NCEA. This environment meant that subjects taught in schools and their instruction varied within and across time, as well as within and across schools.

However, this complexity poses challenges to multilevel models, mainly when the underlying distribution of the outcome variable is binary. Interpretations of model estimations are not straightforward in that the more random effects fitted the more conditions it restricts in inference. It leads to the trade-off between model simplicity and model robustness (as measured in terms of reliability and accuracy).

If the school-specific random effect model had been only used there is the risk of overestimating the fixed effects including the instructional effect of the *LD-Coverage*. The likely reason for this is the intraclass correlation between students and subjects. We must have conservative estimates if we are going to make sound judgements of intervention effects.

However, improvements on model accuracy are mostly due to the inclusion of the student-specific effect alone in comparison to the subject-specific effect even though the intervention of literacy instruction is subject-specific. Moreover, the improved level of model reliability and accuracy are marginal when the subject-specific effect is fitted on top of the student-specific effect.

It offers flexibility in model specifications where on the one hand the inclusion of the student-specific and subject-specific random effects offers more robustness, on the other hand in cases where simplicity is preferred given a threshold of model accuracy, the inclusion of student-specific random effect is enough.

5 CONCLUSION

This thesis had a primary aim: to build robust statistical models with categorical outcome variables to evaluate educational interventions. It is critical for policy and equity reasons to know just how effective a school intervention is and under what conditions. Addressing this aim meant solving two challenges in value-added models in educational intervention research: (a) the lack of uniform valued student outcomes and (b) the complexity of the data structure.

The first practical challenge led to the first research question of this thesis: how to choose between valued student outcomes for your value-added models? Specifically, how to optimise reliability when fitting categorical outcome variables in a continuous outcome model and how to reliably use dichotomised continuous outcome variables in a binary outcome model? The second practical challenge led to the second research question of this thesis: how to specify the model to achieve a good balance in the trade-off between model complexity and model validity, as well as efficiency and accuracy? Specifically, how many underlying random effects should be fitted in a multilevel model with categorical outcomes in order to reliably and effectively determine individual- and group-specific effects.

The findings from the primary school case study chapter showed that with respect to the first research question: (a) categorical outcome variables can be reliably fitted in a continuous model as long as the underly distribution is asymptotically normal (such as stanine scores in standardised tests); and (b) dichotomisation of continuous outcome measures results in a more conservative but more direct estimation of intervention effects.

The findings from the secondary school case study chapter showed that with respect to the second research question: (a) model reliability and accuracy are improved when additional random effects are fitted and (b) the size of improvement varies between random effects where improvements are mostly contributed by the inclusion of the lower-level random effects.

In this chapter, I first summarise the implications in Section 5.1 and the significance of contributions of this thesis in Section 5.2. I then discuss the limitations of the results from the two case studies in Section 5.3.

5.1 Implications

The two case studies, like other studies, collectively show that multilevel model requires careful consideration around the student outcomes and the relevance of the model specifications to the research questions, and the nature of the data structure (Aitkin & Longford, 1986; Aitkin & Zuzovsky, 1994; Goldstein, Burgess, & McConnell, 2007; Larsen & Merlo, 2005).

The choice of valued student outcome in an educational system is often not emphasised in statistical modelling. It is partly because the need to have a consistent valued student outcome is not an issue in systems with a single assessment like the USA or where assessment does not allow the range of choice like NCEA. By contrast, the New Zealand education system is one where there is a plurality of assessments that are nationally recognised at primary level, and where there is flexibility in the national assessment in the secondary level with a choice of assessment measures with varying degrees of measurement scales (e.g., categorical, binary). The flexibility of self-governing allows each school to design their curriculum and assess these within the overall curriculum framework, i.e. NZC. In this environment, a consistent and uniform measure of valued student outcome is required in both cross-sectional and longitudinal intervention studies. This thesis shows that there is a certain degree of adequacy in transforming outcome variables with reduced variance and hence information in examine intervention impact, contrary to what has been suggested in the existing research (Altman & Royston, 2006).

The finding from the primary school case study has a particular practical implication that it shows transformations of outcome variables to a uniformed categorical or binary scale (in cases where outcome variables differ between schools and across data collections) can produce a reliable and accurate estimate of intervention effects. It means that the

evaluation of longitudinal or cross-sectional intervention effects can be reliably assessed even in cases where the assessment of valued student outcomes significantly varies.

The secondary school case study also offers practical implications. I found that controlling for more group-specific effects in complex data structure improves model reliability and accuracy. This finding is consistent with earlier research in school interventions with continuous outcomes (Lai, et al., 2018; Lai, et al., 2014; Jesson, et al., 2018a; Jesson, et al., 2018b; Cervini., 2009a; Timmermans, et al., 2011; Timmermans, et al., 2013). This finding leads to a broader consideration of model specifications in terms of fixed and random effects. By broader consideration, I mean covariates that are traditionally not considered as nesting variables such as subject or department can then be considered for either fixed and or random effects. It is important in the setting of educational research in New Zealand because there is usually clustering that occurs in groups that are not usually clustering variables. For example, year level is not considered a clustering variable rather than a covariate (which is how it is typically dealt with in the literature).

Another practical implication offered by the finding from the secondary school case study is that inclusion of one (or two) lower level random effects can be enough for the desired balance of model accuracy, reliability, and simplicity. It is important as multilevel models using categorical (including binary) outcomes render more model complexity in comparison to models using continuous outcomes (Sommet & Morselli, 2017; Austin & Merlo, 2017). This thesis extends the findings from current research on continuous outcome measures (Lai, et al., 2014; Lai, et al., 2018; Jesson, et al., 2018a; Jesson, et al., 2018b; Cervini., 2009a; Timmermans, et al., 2011; Timmermans, et al., 2013) and shows that in case where model simplicity is required, the inclusion of a lower-level random effect is enough.

5.2 Significance of Contributions

This thesis proposed two solutions to the problems of developing robust quantitative research methods in educational research in New Zealand: a) dichotomisation of valued student outcomes to address the challenge of no uniform valued student outcomes and b) using an integrated perspective to determine the inclusion of random effects by taking considerations of model assumptions, reliability, efficiency and accuracy to address the challenge of the complexity of the underlying data structures. However, the solutions, and thus my contribution extends beyond practical implications (see Section 5.1) in

educational research within New Zealand. Its contribution has both practical and theoretical significance, nationally and internationally. I will outline what I mean below.

5.2.1 Dichotomisation of valued student outcomes

Dichotomisation of valued student outcomes through rigorous alignments to a thorough benchmark can support educational research to a broader extent and enhance the reliability of research findings.

Firstly, dichotomisation offers a robust approach to align different outcome measures from various sources to answer one shared research question. Current studies often use a single assessment measure to approximate students' specific learning outcome despite the lack of direct connections between the assessment measure and intervention (Jesson, et al., 2018a; Jesson, et al., 2018b). For example, Jessen et al. (2018a) used the writing assessment as the valued student outcome to determine effective classroom practice where classroom practices were observed not only in writing but also across reading, mathematics, and science. An alternative approach is to split the sample data by each subject and then integrate the estimated subject-specific effects by meta-analysis. However, such an approach requires a rigorous and nuanced sample size that is substantial to maintain the statistical power and adequate to control the overall significance level (Shein-Shung, Jun, & Hansheng, 2003). Unfortunately, the sample size is often beyond the control of an educational researcher due to ethical reasons. Dichotomisation provides an efficient and effective approach that allows the research question to investigate the sample data both cross-sectionally and longitudinally within the original framework of study design (e.g. sample size).

From a cross-sectional perspective, it allows alignments of outcome measures across not only curriculum areas (e.g. reading, writing, and mathematics) but also other vital competencies that are (traditionally) not always measured quantitatively such as critical thinking, digital awareness, social skills, and well-being. For example, in research aiming to determine effects of a pedagogical intervention on students' learning outcome, the lack of unique student outcome variable often limits the study to a single curriculum, e.g. reading or writing (Lai et al., 2018; Lai et al., 2014; Jesson et al., 2018b; Jesson et al., 2018a). However, the dichotomisation approach renders a unique student learning outcome independent of the taught curriculum in the classrooms. In turn, it extends an intervention study to a multi-strand and multi-curriculum design which allows between-curriculum effects to be measured directly assuming proper classroom randomisation

across curriculums. Take social skills for another example, in research aiming to determine effects of digital usage intervention on students' pro-social skills, the lack of unique student outcome variable in pro-social skills can be solved by dichotomisation of ratings from a set of reliable and valid pro-social questionnaires.

From a longitudinal perspective, it allows alignments of outcome measures across not only year levels or age but also across systematic changes such as changes in government policies and frameworks. For example, in research aiming to determine protective factors of life-long achievement from age 4 to 25, the unique life-long achievement outcome variable can be yield by dichotomisations of valid and reliable achievement outcome measures across different age segmentation, e.g. early childhood, primary school, secondary school, tertiary education, and employment; and tracked longitudinally by aligning to the underlying national framework at the time of data collection. Take changes in policy for another example: one critical issue of the existing educational research programme in New Zealand is the recent abolishment of the New Zealand National Standards. Longitudinal research that stretches across such change of educational framework will suffer a lack of valid student learning outcomes as the valid student outcome has changed since the policy introduction. However, dichotomisation allows the flexibility to align measures under each framework to continue longitudinal research across policy changes beyond the control of research design.

Dichotomisation of valued student outcomes can also be used to cross-validate value-added models to enhance the quality of research findings. One major criticism of the value-added model is the overestimation of intervention effects due to significantly large variations in outcome measures (Briggs, 2008; Briggs & Domingue, 2011). Dichotomisation of valued student outcomes can offer cross-validations under the same model specification. For example, in research aiming to determine effects of literacy intervention on student writing achievement with a norm-referenced continuous measure, we can fit the same model specification to two sets of learning outcomes: one continuous measure and the other dichotomised measure to the norm. The reduction in information allows the dichotomised outcome model to set a qualitative benchmark on how much value was added by the literacy intervention on student writing achievement. The continuous outcome model then can be used to measures the magnitude of such effects. The dichotomised outcome model, together with the continuous model, offers a well-rounded examination of the potential intervention effect.

5.2.2 Determining random effects

This thesis also offers a holistic approach to determine the choice of fixed or random effects in model specifications, taking into considerations of underlying data structures, model assumptions, model accuracy, model efficiency, and model reliability. The trade-off between fixed and random effects renders model specifications to account for data structures with greater complexity and extends the existing theory of value-added models to fit for the original research aims (Goldstein, 1997; Raudenbush, 2004; Timmermans et al., 2011; Foley & Goldstein, 2012; Ray, 2006; Sammons et al., 1993).

Firstly, one critical issue of multilevel models with full realisations of the underlying data structure is the likelihood of a non-convergence in model estimation gets higher as the level of data complexity gets greater (Goldstein, 1997; Raudenbush, 2004). Traditionally, statistical tests were used to determine the choice between fixed and random effects (Goldstein, 1997; Raudenbush, 2004). I argued in this thesis that, with applications to educational research, such full realisation in the model specification is unnecessary from the perspective of the original research aim and a holistic approach other than simple statistical tests are required in determining the choice of fixed and random effects. Such a holistic approach requires the balance of model assumptions, model accuracy, model efficiency, model reliability and model accuracy. For example, in a longitudinal study with multiple nesting groups (e.g. classroom nested within school nested within learning community nested within cities), the choice of random effects ranges from temporal to cross-sectional. This thesis argues that a 3-level model specification is sufficient to yield a model estimation that satisfies the requirements of model accuracy, model reliability and model efficiency with knowing limitations in model assumptions.

This thesis discussed the trade-off between model assumption, accuracy, reliability, and efficiency in multilevel models with both fixed and random effects (De Fraine, Van Damme, Van Landeghem, Opdenakker, & Onghena, 2003; Goldstein et al., 2007; Leckie, 2009; Martínez, 2012; Rasbash, Leckie, & Pillinger, 2010; Timmermans, Snijders, & Bosker, 2013; Troncoso, Pampaka, & Olsen 2015). The two proposed solutions collectively extend the existing theory in value-added modelling in that they render an integrated method that can answer the original research questions with an adequate and consistent outcome measure and a reliable and efficient model specification; and by doing so, address many of the questions and issues raised by other researchers (Goldstein et al., 2007; Leckie, 2009; Martínez, 2012; Rasbash, Leckie, & Pillinger, 2010; Timmermans, Snijders, & Bosker, 2013; Troncoso, Pampaka, & Olsen 2015). It offers a robust process

from model assumption to model interpretation that can be reliably repeated in applications to evaluate intervention success.

The practical and theoretical contributions of this thesis also extend beyond the field of educational research to biostatistics, social and political science, public health, econometrics, and psychology where a) the underlying outcomes or responses or endpoints are categorical and b) the intervention follows a randomised-control design (Goldstein et al., 2007, Gelman & Hills, 2007). For example, in a randomised-control political science study across multi-centres (e.g. Europe, Japan, USA and China) where the underlying data structures were complex due to the different study timeframes within and across each centre, method used in Chapter 4 can be applied to determine the level of fixed and random effects in the model specification.

5.3 Limitations

The results of the two case studies are based on school with relatively homogenous populations in terms of ethnicity and socioeconomic status. Both are primarily low decile with over-represented Māori and Pasifika students, and in both cases, the interventions are organised in clusters of schools and have evidence of previous intervention successes. The data sets were chosen for the modelling purposes of this thesis due to its availability and reliability. I had access to these data sets (for confidentiality reasons, such data sets are often not available for secondary data analysis). These data sets had adequate sample sizes for modelling and are collected reliably. However, such homogeneity has some limitations. It reduces the intra-subject and intra-school correlations that may lead to an under-estimated subject-specific effect.

Previous intervention success, although it is not a guarantee for current intervention effect, could also have increased the likelihood of detecting effects even in cases where variability is reduced (e.g. transformations of outcome variables). While in future, it would be useful to have a larger and less homogenous data set. In reality, data sets for interventions will have some of the limitations identified in this thesis.

Any statistical package has inbuilt limitations and assumptions. These were selected because they were recommended in the literature (Sommet & Morselli, 2017; Austin & Merlo, 2017), and most importantly, they are open-sourced. These packages used in this thesis are well applied in studies of multilevel models. However, more advanced packages were developed with improved efficiency and accuracy in recent months during the writing of this thesis (Stan Development Team, 2018). They could have influenced the

results in models with the logit link where model estimations rely on adequate convergence. However, considering the widespread and published recommended use, they were appropriate for the thesis. Considering all packages have limitations, it is more important to explain these limitations explicitly and how they could have influenced the results.

Last but not least, the underlying study must follow a well-designed randomised-control framework or a quasi-experimental study with a well-established comparable baseline in order to make causal inference based on the study results (Gelman & Hills, 2007).

In conclusion, this thesis shows first that it is possible to model data longitudinally and cross-sectional in the absence of a unique valued student outcome via straightforward transformations of the outcome variables such as alignment to the New Zealand curriculum levels across year levels. From a practical perspective, this provides a straightforward and cost-effective approach to analysing existing data sources from different assessments to uncover the effectiveness of government interventions. This approach does not require a new dataset but can be a secondary analysis of existing data sets.

This thesis also shows that fitting random effect closer to the underlying data structure improves intervention evaluation. However, in considering trade-offs between model simplicity and model effectiveness (in terms of reliability and accuracy), it is sufficient to model the lower-level random effects.

6 BIBLIOGRAPHY

AERA. (2015). Statement on Use of Value-Added Models (VAM) for the Evaluation of Educators and Educator Preparation Programs. (2015). *Educational Researcher*, 44(8), 448–452.

Agresti, A. (2013) *Categorical Data Analysis*. 3rd Edition, John Wiley & Sons Inc., Hoboken.

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of Royal Statistical Society*, 149, 1, 1-43.

Aitkin, M., & Zuzovsky, R. (1994). Multilevel interaction models and their use in the analysis of large-scale school effectiveness studies. *School Effectiveness and School Improvement*, 5, 1, 45-73.

Aitkin, M., Anderson, D., and Hinde, J. (1981). Statistical modeling of data on teaching styles (with discussion). *Journal of the Royal Statistical Society*, A 144, 148-161.

Albert, J. H. (1988). Computational methods using a Bayesian hierarchical generalized linear model. *Journal of the American Statistical Association*, 83, 404, 1037-1044.

Armitage, C. J. (2008). Cognitive and affective predictors of academic achievement in school children. *British Journal of Psychology*, 99, 57-74. doi: 10.1348/000712607X181313

Antony, F. (2002). Teaching groups as foci for evaluating performance in cost effectiveness of GCE advanced level provision: Some practical methodological innovations. *School Effectiveness and School Improvement*, 13, 2, 225-246.

Arbuthnot, J., & Wayner, M. (1982). Minority influence: Effects of size, conversion, and sex. *Journal of Psychology*, 111, 2.

Arnold, C. L. (1995). *Using HLM and NAEP Data to explore school correlates of 1990 mathematics and geometry achievement in grades 4, 8, and 12: Methodology and results*.

A Technical Report. U. S. Department of Education, Office of Educational Research and Improvement, Washington, D. C.

Arnold, C. L., & Kaufman, P. D. (1992). *School effects on educational achievement in mathematics and science: 1985-86.* A Technical Report. National Assessment of Educational Progress, Washington, D. C.

Arnold, C. A. (1993). *Using HLM with NAEP*. Unpublished Paper Presented at the Advanced Studies Seminar on the Use of NAEP Data for Research and Policy Discussion, Washington D. C.

Austin, P. C., & Merlo, J. (2017) Intermediate and advanced topics in multilevel logistic regression analysis. *Statist. Med*., 36: 3257–3277. doi: 10.1002/sim.7336.

Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. BMJ (Clinical research ed.), 332(7549), 1080.

Azibo, D. A. Y. (1983). Perceived attractiveness and the Black personality. *The Western Journal of Black Studies*, 7, 4, 229-238.

Baker, E., Barton, P. E., Haertel, E,. & F.H, (2010). Problems with the use of student test scores to evaluate teachers. Economic Policy Institute.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, 29(1), 37-65.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Bell, B.A., Ene, M., Smiley, W., & Schoeneberger, J.A. (2013). *A Multilevel Model Primer Using SAS*.

Bishop, R., Berryman, M., & Richardson, C. (2003). *Te kotahitanga: The experiences of Year 9 and 10 Māori students in mainstream classrooms*. Wellington: Ministry of Education.

Blaiklock, K. (2004). A critique of running records of children‟s oral reading. *New Zealand Journal of Educational Studies*, 39, 241-253.

Boe, E. E., Barkanic, G., & Leow, C. S. (1999). *Retention and Attrition of Teachers at the School Level: National Trends and Predictors*. A Technical Report. Center for Research and Evalustion in Social Policy, Graduate School of Education, University of Pennsylvania.

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton: Educational Testing Service. Accessed 27 February 2008.

Braun, H. I., Jones, D. H., Rubin, D. B., & Thayer, D. T. (1983). Empirical Bayes estimation of coefficient in general linear model from data of different rank. *Psychometrika*, 289(2), 171-181.

Braun, H., Chudowsky, N., & Keonig, J. (2010). *Getting Value Out of Value-Added*. Washington, DC: National Academics Press.

Breslow, N., & Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.

Brody, G. H., Ge, X., Conger R., Gibbons, F. X., Murry, V. M., Gerrard, M., & Simons,

R. L. (2001). The influence of neighborhood disadvantage, collective socialization, and parenting on African American children's affiliation with deviant peers. *Child Development*, 72, 4, 1231-1246.

Bliese, P. D. (1998). Group size, ICC values, and group-level correlations: A simulation. *Organizational Research Methods*, 1 , 355–373.

Briggs, D. C. (2008) Synthesizing causal inferences. *Educational Researcher*, 37(1), 15-22

Briggs, D. C. (2012). Making value-added inferences from large-scale assessments. *Improving large-scale assessment in education: Theory, issues and practice*, 186-206.

Briggs, D. C., & Domingue, B. D. (2011). Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District Teachers by The Los Angles Times. National Education Policy Centre. http://nepc.colorado.edu/publication/due-deligence.

Breslow, N. E. (1996). Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association*, 91(433), 14-28.

Bryk, A. S., & Thum, Y. M. (1989). The effects of high school organization on dropping out: An exploratory investigation. *American Educational Research Journal*, 26, 3, 353-383.

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education*, 97, 1, 65-108.

Bond, L. A. (1996). Norm-and criterion-referenced testing.

Buckendahl, C. W., & Ferdous, A. (2002). *Fourth Grade State Writing Assessment Standard Setting Study*. A Technical Report. Buros Institute for Assessment Consultation and Outreach. University of Nebraska-Lincoln.

Buddin, R. (2010). How effective are Los Angeles elementary teachers and schools?.

Buonaccorsi, J. P. (2010). *Measurement error: models, methods, and applications*. Chapman and Hall/CRC.

Burton, B. (1993). *Some observations on the effect of centering on the results obtained from hierarchical linear modeling*. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Cameron, A. C., & Trivedi, P. K. (2010). *Microeconometrics using stata (Vol. 2)*. College Station, TX: Stata press.

Cervini, R. (2009a). Class, school, municipal, and state effects on mathematics achievement in Argentina: A multilevel analysis. *School Effectiveness and School Improvement*, 20, 319–340.

Chen, C., Campbell, V. C., & Suleiman, A. (2001). *Predicting Student Performances at a Minority Professional School*. Paper presented at the Annual Meeting of the Association for Institutional Research, 41st, Long Beach, CA.

Cheong, Y. F., Fotiu, R. P., & Raudenbush, S. W. (2001). Efficiency and robustness of alternative estimators for two-and three-level models: The case of NAEP. *Journal of Educational and Behavioral Statistics*, 26, 4, 411-429.

Cizek, G. J., & Fitzgerald, S. M. (1999). Methods, plainly speaking: an introduction to logistic regression. *Measurement and evaluation in counseling and development*, 31, 223-245.

Clay, M. M. (2000). *Running Records for Classroom Teachers*. Auckland: Marie Clay Trust.

Cohen, M. P. (2000). Note on the odds ratio and probability ratio. *Journal of Educational and Behavioral Statistics*, 25, 2, 249-252.

Darr, C., McDowall, S., Ferral, H., Twist, J., and Watson, V. (2008). *Progressive Achievement Test: Reading—teacher manual (2nd ed.)*. Wellington: New Zealand Council for Educational Research.

Darr, C., Neill, A., Stephanou, A., & Ferral, H. (2007). *PAT Progressive Achievement Test: Mathematics (2nd ed., rev.)*. Wellington: New Zealand Council for Educational Research.

Davidian, M., & Gallant, R. (1993).The Nonlinear Mixed Effects Model with a Smooth Random Effects Density. *Biometrika*, 80, 474-488.

De Fraine, B., Van Damme, J., Van Landeghem, G., Opdenakker, M.-C., & Onghena, P. (2003). The effect of schools and classes on language achievement. *British Educational Research Journal*, 29, 841–859

Deng, B. (1992). *A multilevel analysis of classroom climate effects on mathematics achievement of fourth-grade students*. Unpublished doctoral dissertation. Memphis State University, TN.

Dobson, A. 2002. *An Introduction to Generalized Linear Models. Chapman & Hall/CRC*: Boca Raton, FL.

Draper, D. (1995). Inference and hierarchical modeling in social sciences. *Journal of Educational and Behavioral Statistics*, 20, 2, 115-147.

Du, Y. & Heistad, D. (1999). *School performance accountability in Minneapolis Public Schools*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Elley, W., Ferral, H., & Watson, V. (2011). *STAR Supplementary Test of Achievement in Reading (2nd ed.)*. Wellington: New Zealand Council for Educational Research.

Felch, J., Song, J., & Smith, D. (2010). Who's teaching LA's kids? A Times analysis, using data largely ignored by LAUSD, looks at which educators help students learn, and which hold them back. Los Angeles Times.

Ferguson, J. M. (1999). The effect of year-round school on student achievement in mathematics. *The Educational Forum*, 64, 82-87.

Ferrão, M. E., & Goldstein, H. (2009). Adjusting for measurement error in the value added model: Evidence from Portugal. *Quality & Quantity*, 43, 951–963.

Finch, W. H., & Cassady, J. C. (2014). Use of the probit model to estimate school performance in student attainment of achievement testing standards. *Educational Assessment, Evaluation and Accountability*. doi:10.1007/s11092-013-9186-6

Fisher, F. W. (2001). *Multinomial logistic regression modeling in vocational rehabilitation*. Unpublished doctoral dissertation. Florida State University, FL.

Fitzmaurice, G. M., & Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, 80, 1, 141-151.

Fleiss, J. (1986). *The Design and Analysis of Clinical Experiments*. Wiley, New York.

Foley, B., & Goldstein, H. (2012*). Measuring success. League tables in the public sector*. London, UK: The British Academy.

Fuller, B. (1987). What school factors raise achievement in the third world?. *Review of educational research*, 57(3), 255-292.

Gardiner, Joseph C.; Luo, Zhehui; Roman, Lee Anne (2009). "Fixed effects, random effects and GEE: What are the differences?". *Statistics in Medicine*. 28: 221–239. doi:10.1002/sim.3478.

Gelman, A. (2004). *Analysis of variance—why it is more important than eve*r. The Annals of Statistics.

Gelman, A. (2006). Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics*, 48(3), 432-435.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press.

Gibson, A., & Asthana, S. (1998). School performance, school effectiveness and the 1997 White Paper. *Oxford Review of Education*, 24(2), 195-210.

Gill, J. (2003). Hierarchical linear models. In Kimberly Kempf-Leonard (Ed.), *Encyclopedia of social measurement*. New York: Academic Press.

Glazerman, S., Loeb, S., Goldhaber, D. D., Raudenbush, S., Whitehust, G. J., & Policy, B.I.B.C.E. (2010). *Evaluating teachers: The important role of value-added. New York*. Brown Center on Education Policy at Brookings.

Goldstein, H. (1995). *Multilevel statistical models* (second edition). New York: John Wiley.

Goldstein, H., & Rabash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society*, 159 (part 3), 505-513.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73, 1, 43-56.

Goldstein, H. (1987a). *Multilevel models in educational and social research*. London: Griffin.

Goldstein, H. (1987b). Multilevel covariance component models. *Biometrika* 74, 430-431.

Goldstein, H. (1997). Methods in school effectiveness research. School Effectiveness and School Improvement, 8, 369–395

Goldstein, H. (2001). Using pupil performance data for judging schools and teachers: Scope and limitations. *British Educational Research Journal*, 27, 433–442.

Goldstein, H., Burgess, S., & McConnell, B. (2007). Modelling the effect of pupil mobility on school differences in educational achievement. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170, 941–954.

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). Chichester, UK: John Wiley and Sons.

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society*, A. 159, 385-443.

Goldhaber, D., & Hansen, M. (2010). Using performance on the job to inform teacher tenure decisions. *American Economic Review*, 100(2), 250-255.

Gordon, R., Kane, T., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job*. The Brookings Institution.

Greene, W. 2000. *Econometric Analysis. Fourth Edition*. Prentice Hall: Upper Saddle River, NJ.

Guo, G., & Zhao, H. (2000). Multilevel modeling for binary data. *Annual Review of Sociology*, 26, 441-462.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of 800+ meta-analyses on achievement*. London, England: Routledge.

Hannan, P. J., & Murray, D. M. (1996). Gauss or Bernoulli? A Monte Carlo comparison of the performance of the linear mixed-model and logistic mixed-model analyses in simulated community trials with a dichotomous outcome variable at the individual level. *Evaluation Review*, 20, 3, 338-352.

Hanushek, E. A. (1971). Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro-Data, American Economic Review, 61(2), pp. 280-288;

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, 100(2), 267-271.

Hargrove, L., L. & Mao, M. X. (1997). *Three-level HLM modeling of academic and contextual variables related to SAT scores in Texas*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Hanushek, E. A. (1992). The Trade-off Between Child Quantity and Quality, *Journal of Political Economy*, 100(1), pp. 84-117.

Harris, J. (2007). Government tweaks NCEA. *New Zealand Education Review*, 12(20),

Hartig, F. (2019). DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models. R package version 0.2.3. https://CRAN.R-project.org/package=DHARMa

Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? An examination of the statistical properties and policy alternatives. *Education Finance and Policy, 4(4)*, 319-350.

Hartig, F. (2019). DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models. R package version 0.2.3. https://CRAN.R-project.org/package=DHARMa

Heck, R. H. (2000). Examining the Impact of School Quality on School Outcomes and Improvement: A Value-Added Approach. *Educational Administration Quarterly*, 36,

513-552. Retrieved 24 November, 2009, Retrieved from http://eaq.sagepub.com/cgi/content/abstract/36/4/513

Heistad, D. (1999). *Teachers who beat the odds: Value-added reading instruction in Minneapolis 2nd grade classrooms.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada, April 19-23.

Hill, H. C., Kapitula, L., & Umland, K. (2010). A validity argument to evaluating teacher value-added scores. *American Educational Research Journal*, 48(3), 794-831.

Hill, P. W., & Goldstein, H. (1998). Multilevel Modeling of Educational Data With CrossClassification and Missing Identification for Units. *Journal of Educational and Behavioral Statistics*, 23, 117-128. Retrieved 20 November, 2009, Retrieved from http://jeb.sagepub.com/cgi/content/abstract/23/2/117

Hodis, F.A., Meyer, L.H., McClure, J., Weir, K., & Walkey, F. (2011). A Longitudinal Investigation of Motivation and Secondary School Achievement Using Growth Mixture Modelling. *Journal of Educational Psychology*, Vol. 103, No. 2, 312-323.

Hoeksma, J. B., & Knol, D. L. (2001). Testing predictive developmental hypotheses. *Multivariate Behavioral Research*, 36(2), 227-248.

Hoffmann, J. P. (2004). *Generalized linear models: an applied approach*. Pearson: Boston.

Hofmann, D. A. (1997). An overview of the logic and rationale of hierarchical linear models. *Journal of Management*, 23, 723-744.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York: John Wiley.

Hox, J. (2002). *Multilevel Analysis Techniques and Applications*. New Jersey: Lawrence Erlbaum.

Hox, J. (2010). *Multilevel analysis: Techniques and applications. Quantitative methodology series (2nd ed.).* New York, NY: Routledge.

Hsiao, C. (2003). *"Fixed-effects models". Analysis of Panel Data (2nd ed.)*. New York: Cambridge University Press. pp. 95–103. ISBN 0-521-52271-4.

Ishii, J., & Rivkin, S. G. (2009). Impediments to the estimation of teacher value added. *Education Finance and Policy*, *4*(4), 520-536.

James, D. L., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *The Journal of Applied Psychology*, 69, 85–98.

Jesson, R., McNaughton, S., Rosedale, N., Zhu, T., & Cockle, V. (2018). A mixed-methods study to identify effective practices in the teaching of writing in a digital learning

environment in low income schools. *Computers & Education*, 119, 14-30. 10.1016/j.compedu.2017.12.005

Jesson, R., McNaughton, S., Wilson, A., Zhu, T., & Cockle, V. (2018). Improving Achievement Using Digital Pedagogy: Impact of a Research Practice Partnership in New Zealand. *Journal of Research on Technology in Education*, 50 (3), 183-199. 10.1080/15391523.2018.1436012

Jimenez, J. D. & Salas-Velasco, M. (2000). Modeling educational choices: A binomial logit model applied to the demand for higher education. *Higher Education*, 40, 3, 293-31.

Jonson-Reid, M., Williams, J. H., & Webster, D. (2001). Severe emotional disturbances and violent offending among incarcerated adolescents. *Social Work Research*, 25, 4, 213-222.

Kane, T. J., & Staiger, D. O. (2002). The promise and pitfall of using imprecise school accountability measures. Journal of Economic Perspectives, 16(4), 91-114.

Kane, T., & Staiger, D. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. *NBER working paper*. Retrieved from http://www.nber.org/papers/w14607

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy*, 6(1), 18-42.

Kang, S. J. (1993). *A covariance component model with two-way crossed random effects and ML estimation via the EM algorithm*. Unpublished doctoral dissertation, University of Michigan, MI.

Kasim, R. M. & Raudenbush, S. W. (1998). Application of Gibbs sampling to nested variance components models with heterogeneous within-group variance. *Journal of Educational and Behavioral Statistics*, 23, 2, 93-116.

Kelly, A., & Downey, C. (2010). Value-added measures for schools in England: looking inside the 'black box'of complex metrics. *Educational Assessment, Evaluation and Accountability*, 22(3), 181-198.

Kevan, R. (1997). *A demonstration of the HLM growth modeling technique applied to native American population*. Unpublished doctoral dissertation. Florida State University, Department of Educational Research.

Kirk, R. (1995). *Experimental Design: Procedures for the Behavioral Sciences* (third edition). Pacific Grove, CA: Brooks/Cole, 1995.

Kreft, I. G. G., Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical models. *Multivariate Behavioral Research*, 30(1), 1-21.

Kreft, I. G. G. (1995). *The Effects of Centering in Multilevel Analysis: Is the Public School the Loser or the Winner?* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Lai, M. K., McNaughton, S., Amituanai-Toloa, M., Turner, R., & Hsiao, S. (2009). Sustained acceleration of achievement in reading comprehension: The New Zealand experience. *Reading Research Quarterly*, 44(1), 30-56. doi:10.1598/RRQ.23.3.2

Lai, M. K., McNaughton, S., & Zhu, T. (2018). *Instructional Practices in Data Discussions and their Relationship with Achievement.* Paper presented at the ICSEI conference, Singapore, January 2018.

Lai, M. K., Wilson, A., McNaughton, S., & Hsiao, S. (2014). Improving achievement in secondary schools: Impact of a literacy project on reading comprehension and secondary school qualifications. *Reading Research Quarterly*, 49(3), 305. doi:DOI: 10.1002/rrq.73

Laird, N. M., & Ware, H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963-974.

Larsen, K., & Merlo, J. (2005) Appropriate assessment of neighborhood effects on individual health: integrating random and fixed effects in multilevel logistic regression. *American Journal of Epidemiology*, 161(1):81–88.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, 11 , 815–852.

Leckie, G. (2009). The complexity of school and neighbourhood effects and movements of pupils on school differences in models of educational achievement. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172, 537–554.

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007a) Bayesian Methods for Scalable Multivariate Value-Added Assessment. *Journal of Educational and Behavioral Statistics*, 32(2). 125-150.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage Publications.

Longford, N. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced models with nested random effects, *Biometrka*, 74, 4, 817-827.

Ma, X. (1999). Dropping out of advanced mathematics: the effects of parental involvement. *Teachers College Record*, 101, 1, 60-81.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. Methodology: European *Journal of Research Methods for the Behavioral and Social Sciences*, 1(3), 86-92.

Madjar, I., & McKinley, E. (2011). *Understanding NCEA: A relatively short and very useful guide for secondary school students and their parents*. NZCER Press, Wellington.

Martínez, J. F. (2012). Consequences of omitting the classroom in multilevel models of schooling: An illustration using opportunity to learn and reading achievement. *School Effectiveness and School Improvement*, 23, 305–326.

Mason, W. M., Wong, G. M., & Entwistle, B. (1983). Contextual analysis through the multilevel linear model. In S. Leinhardt (Ed.), *Sociological Methodology* (pp 72-103). San Franscisco: Jossey-Bass.

McCaffrey, D. F. (2003). Evaluating value-added models for teacher accountability (Vol. 158). Santa Monica: Rand Corporation.

McCaffrey, D. F., Lockwood, J., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, 29(1), 67.

McCaffrey, D. F., Han, B., & Lockwood, J. (2009). Turning student test score into teacher compensation system,

McCaul, E. J., Donaldson Jr., G., Coladarci, T., & Davis, W. E. (1992). Consequences of dropping out of school: Findings from high school and beyond. *Journal of Educational Research*, 85, 4, 198-207

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (second edition). London: Chapman & Hall.

McLaughlin, D., & Drori, G. (2000). *School-level correlates of academic achievement: Student assessment scores in SASS public schools*. A Technical Report. National Center for Educational Statistics, Washington, D. C.

Menard, S. (1995). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage.

Meyer, L. H., McClure, J., Walkey, F., Weir, K. F., & McKenzie, L. (2009). Secondary student motivation orientations and standards-based achievement outcomes. *British Journal of Educational Psychology*, 79(2), 273-293.

Meyer, L. H., Weir, K. F., McClure, J., Walkey, F., & McKenzie, L. (2009). *Motivation and Achievement at Secondary School: The Relationship Between NCEA Design and Student Motivation and Achievement: a Three Year Follow-up*. Victoria University of Wellington.

Ministry of Education. (n.d.). Assessment resources maps.

Retrieved from: http://assessment.tki.org.nz/Assessment-tools-resources/Assessment-resources-maps

Ministry of Education. (2000). *Using Running Records*. Wellington: Learning Media.

Ministry of Education. (2007). *The New Zealand curriculum*. Wellington: Learning Media.

Ministry of Education. (2009). *The New Zealand curriculum reading and writing standards for years 1–8*. Wellington: Learning Media.

Ministry of Education. (2012). *New Zealand schools ngā kura o Aotearoa 2011: A report on the compulsory schools sector in New Zealand—2011*. Wellington: Author.

Ministry of Education. (2013). Research into the Implementation of the Secondary Literacy Project (SLP) in Schools. Retrieved from http://www.educationcounts.govt.nz/publications/series/Secondary_Literacy

Morgan, G. A. (2000). Quasi-Experimental Designs. *Journal of the American Academy of Child & Adolescent Psychiatry*. 39. pp. 794–796. doi:10.1097/00004583-200006000-00020.

Morganstein, D., & Wasserstein, R. (2014) ASA Statement on Value-Added Models, Statistics and Public Policy, 1:1, 108-110, DOI: 10.1080/2330443X.2014.956906

Muthen, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, 28, 4, 338-54.

Nakagawa, S., Johnson, P.C.D., Schielzeth, H. (2017) The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J. R. Soc. Interface* 14: 20170213.

Nakagawa, S., Schielzeth, H. (2013) A general and simple method for obtaining R^2 from Generalized Linear Mixed-effects Models. *Methods in Ecology and Evolution* 4: 133–142

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *education policy analysis archives*, 18, 23.

New Zealand Council For Educational Research. (2012). *e-asTTle writing (revised) Manual*. Prepared for Ministry of Education.

New Zealand Qualifications Authority. (2009). NCEA web-based information. Retrieved from http://www.nzqa.govt.nz

New Zealand Qualifications Authority. (2014). *NZQA: Annual Report on NCEA and New Zealand Scholarship Data and Statistics (2013)*, NZQA, Wellington.

Onis, M. D. (2000). Measuring nutritional status in relation to mortality. *Bulletin of the World Health Organization*, 78 (10), 1271-74.

Patrick, W. J. (2000). *Estimating first-year student attrition rates: An application of multilevel modeling using categorical variables.* Paper Presented at the 40th Annual Forum of The Association for Institutional Research, Cincinnati, Ohio.

Peng, C. J, So, T. H., Stage, F. K., & John, E. P. (2002). The use and interpretation of logistic regression in higher education journals: 1988-1999. *Research in higher education*, 43, 3, 259-293.

Phillips, G. W. & Adcock, E. P. (1997). *Measuring school effects with hierarchical linear modeling: Data handling and modeling issues*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Pike, G. R., & Saupe, J. L. (2002). Does high school matter? An analysis of three methods of predicting first-year grades. *Research in Higher Education*, 43, 2, 187-207.

Pituch, K. (1997). *Ranking school effects with a two-level hierarchical linear model: Illustration and problems with current practice*. Unpublished doctoral dissertation. Florida State University, FL.

PM Benchmarks. (n.d.) Retrieved from http://assessment.tki.org.nz/Assessment-tools-resources/Assessment-tool-selector/Browse-assessment-tools/English/Reading/PM-Benchmarks/(back_to_results)/Assessment-tools-resources/Assessment-tool-selector/Browse-assessment-tools

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rabe-Hesketh, S. and Skrondal, A. (2012). Multilevel and Longitudinal Modeling Using Stata (Third Edition). College Station, TX: Stata Press. Volume 2.

Rabash, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and crossclassified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics*, 19, 4, 337-350.

Rachman-Moore, D., & Wolfe, R. G. (1984). Robust analysis of a nonlinear model for multilevel educational survey data. *Journal of Educational Statistics*, 9, 4, 277-293.

Rasbash, J., Leckie, G., & Pillinger, R. (2010). Children's educational progress: Partitioning family, school and area effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173, 657–682.

Raudenbush, S. W. & Bryk, A. (2002). *Hierarchical linear models in social and behavioral research: Applications and data analysis methods* (second edition). Newbury Park, CA: Sage.

Raudenbush, S. W., Bryk, A., Cheong, Y. F., & Congdon, R. (2001). *HLM 5: Hierarchical linear and nonlinear modeling*. Chicago, IL: Scientific Software International.

Raudenbush, S. W., Bryk, A., & Congdon, R. (2000). *Hierarchical linear and nonlinear modeling (HLM)*. Chicago, IL: Scientific Software International.

Raudenbush, S. W., Bryk, A., & Ponsiciak. S. (2003). School accountability. Paper presented at annual meeting of American Educational Research Association, Chicago, Il.

Raudenbush, S. W., Fotiu, R., & Cheong, Y. F. (1999). Synthesizing results from the Trial State Assessment. *Journal of Educational and Behavioral Statistics*, 24(4), 413-438.

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, 20, 4, 307-335.

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with application in cross-sectional and longitudinal research. *Journal of Educational Statistics*, 18, 321-349.

Raudenbush, S. W. (1988). Educational applications of hierarchical linear model: A review. *Journal of Educational Statistics*, 13, 2, 85-116.

Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, 59, 1-17.

Raudenbush, S. W. Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods, second edition. Newbury Park*, CA: Sage.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy*, 4(4), 492-519.

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, 29, 121–129

Ray, A. (2006). School value added measures in England. Retrieved from https://education.gov.uk/publications/eOrderingDownload/RW85.pdf

Rindskopf, D. (2002). Infinite parameter estimates in logistic regression: Opportunities, not problems. *Journal of Educational and Behavioral Statistics*, 27, 2, 147-161.

Robinson, G. K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. *Statist. Sci. 6*, no. 1, 15--32. doi:10.1214/ss/1177011926. https://projecteuclid.org/euclid.ss/1177011926

Robinson, V., Bendikson, L., McNaughton, S., Wilson, A., & Zhu, T. (2017). Joining the dots: The challenge of creating coherent school improvement. *Teachers College Record*, 119 (8), 1-44.

Rothstein, J. (2010) Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, 125(1), 175-214.

Rowan, B., Correnti, R., & Miller, R. J. (2002). *What large-scale, survey research tells us about teacher effects on student achievement: Insights from the prospects study of elementary schools*. A Technical Report. University of Pennsylvania, Graduate School of Education.

Rumberger, R. W. (1995). Dropping out of middle school: A multilevel analysis of students and schools. *American Educational Research Journal*, 32, 3, 583-625.

Sadler, P. M., & Hammerman, J. K. (1999). Employing quantitative models of a qualitative admissions process: Uncovering hidden rules, saving time, and reducing bias. *College and University*, 74, 3, 8-24.

Saint-Blancat, C. (1995). The effect of minority group vitality upon its sociopsychological behavior and strategies: An empirical study on the Valdotan minority - Part II. *Journal of Multilingual and Multicultural Development*, 6, 1, 31-44.

Sammons, P., Nuttall, D., & Cuttance, P. (1993). Differential school effectiveness: Results from a reanalysis of the inner London education authority's junior school project data. *British Educational Research Journal*, 19, 381–405

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee ValueAdded Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, 12(3), 247–256.

Sanders, W. L., Wright, S. P., Rivers, J. C., & Leandro, J. G. (2009). A response to criticisms of SAS EVAAS. Cary: SAS Institute Inc.

Saxe, G. B, Gearhart, M., & Seltzer, M. (1999). Relations between classroom practices and student learning in the classroom practices and student learning in the domain. *Cognition and Instruction*, 17(1), 1-24.

Schaffer, W. D., Yen, S. J., & Rahman, T. (2000). School effects indices: Stability of one- and two-level formulations. *The Journal of Experimental Education*, 68, 3, 239-250.

Scherrer, J. (2011). Measuring teaching using value-added modeling: The imperfect panacea. *NASSP Bulletin*, 95(2), 122- 140. doi:10.1177/0192636511410052

Schumacker, R. E., & Bembry, K. (1995). *Centering effects in HLM level-1 predictor variables.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Scott, D. (2008). *How does achievement at school affect achievement in tertiary education?* Wellington, New Zealand: Ministry of Education.

Seber, G. 1980. *Linear regression analysis*. John Wiley and Sons: New York

Searle, S.R., Casella, G. and McCulloch, C.E. (1992). *Variance Components*. Wiley, New York.

http://dx.doi.org/10.1002/9780470316856

Seltzer, M. H. (1993). Sensitivity analysis in fixed effects in the hierarchical model: A Gibbs sampling approach. *Journal of Educational Statistics*, 18, 3, 207-235.

Seltzer, M., Choi, K., & Thum, Y. M. (2002). *Latent variable modeling in the hierarchical modeling framework: Exploring initial status∆treatment interactions in longitudinal studies*. A Technical Report. Graduate School of Education and Information Studies, University of California at Los Angels.

Shay, S. A., & Gomez, J. J. (2002). *Privatization in education: A growth curve analysis of achievement*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Shein-Shung C, Jun S, Hansheng W. (2003). *Sample Size Calculation in Clinical Trial*. New York: Marcel Dekker Inc; 2003. Chapter 1, page 11, section 1.2.3.

Shulruf, B., Hattie, J., & Tumen, S. (2008). Individual and school factors affecting students' participation and success in higher education. *Higher Education*, 56(5), 613-632.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. Educational Researcher, 15(2), 4-14. doi:10.3102/0013189X015002004

Simon, M., Ercikan, K., & Rousseau, M (Eds.) (2012) *Improving Large-Scale Assessment in Education: Theory, Issues and Practice*. London: Routledge

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, 23(4), 323, 355.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY, US: Oxford University Press.

Slack, J. B., & St. John, E. P. (1999). *A practical model for measuring the effect of school reform on the reading achievement of non-transient learners*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Snijders, T. A., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, 18, 3, 237-259.

Sommet, N., & Morselli, D. (2017). Keep calm and learn multilevel logistic modeling: A simplified three-step procedure using Stata, R, Mplus, and SPSS. *International Review of Social Psychology*, 30(1). Article ID 203-218.

Sommet, N., & Morselli, D. (2017). Keep calm and learn multilevel logistic modeling: A simplified three-step procedure using Stata, R, Mplus, and SPSS. *International Review of Social Psychology*, 30(1). Article ID 203-218.

Stage, S. A. (2001). Program evaluation using hierarchical linear modeling with curriculum-based measurement reading probes. *School Psychology Quarterly*, 16, 1, 91-112.

Stan Development Team (2018). RStan: the R interface to Stan. R package version 2.18.2. http://mc-stan.org/.

Strand, S. (1998). A 'value added' analysis of the 1996 primary school performance tables. *Educational Research* -London then Slough-, 40, 123-138. Retrieved 23 November, 2009, from http://www2.warwick.ac.uk/fac/soc/cedar/staff/stevestrand/strand1998er.pdf

Strathdee, R., & Boustead, T. (2005). Measuring “value added” in New Zealand schools. *New Zealand Annual Review of Education*, 14, 59-76.

Subedi, B. R. (2003). Factors influencing high school achievement in Nepal. *International Education Journa*l, 4, 2, 98-107.

Sullivan, L. M., Dukes, K. A., & Losina, E. (1999). Tutorial in biostatistics: An introduction to hierarchical linear modeling. *Statistics in Medicine*, 18, 855-888.

Susman E. , Murphy, J., Zerbe, G., & Jones, R. (1998). Using a nonlinear mixed model to evaluate three models of human stature. *Growth, Development & Aging*, 62, 4.

Tate, R. (2004). Interpreting hierarchical linear and hierarchical generalized linear models with slopes as outcomes. *The Journal of Experimental Education*, 73, 71-95.

Tate, R. (2000). *A study of value-added models for school assessment*. A Technical Report. Florida Department of Education and Florida State University, Tallahassee, FL.

Tate, R. L., & Wongbundhit, Y. (1983). Random versus nonrandom coefficient models for multilevel analysis. *Journal of Educational Statistics*, 8, 2, 103-120.

The University of Auckland (n.d.). Retrieved from https://www.auckland.ac.nz/en/about/admission-and-enrolment/ae-undergraduatestudents/ae-entry-requirements/ae-domestic-students/ae-national-certificate-of-educationalachievement.html

Thomas, S., Sammons, P., Mortimore, P., & Smees, R. (1997). Differential Secondary School Effectiveness: Comparing the Performance of Different Pupil Groups. *British*

Educational Research Journal, 23, 451-469. Retrieved 20 November, 2009, from http://www.jstor.org/stable/1502081

Thum, Y. M. (1997). Hierarchical linear model for multilevel outcomes. *Journal of Educational and Behavioral Statistics*, 22, 1, 77-108.

Timmermans, A. C., Doolaard, S., & De Wolf, I. (2011). Conceptual and empirical differences among various value-added models for accountability. *School Effectiveness and School Improvement*, 22, 393–413

Timmermans, A. C., Snijders, T. A. B., & Bosker, R. J. (2013). In search of value added in the case of complex school effects. *Educational and Psychological Measurement*, 73, 210–228.

Troncoso, P., Pampaka, M., Olsen, W. (2017) Beyond traditional school value-added models: a multilevel analysis of complex school effects in Chile, *School Effectiveness and School Improvement* 27 (3), 293-314

Tuuta, M., Bradnam, L., Hynds, A., & Higgins, J. with Broughton, R. (2004). *Evaluation of Te Kauhua Māori in mainstream pilot project*. Wellington: Minister of Education.

Van Hoef, J., & Boveng, P. (2007). Quasi-poisson vs. negative binomial regression: how should we model overdispersion count data? Ecology. 88: 2766-2772.

Webster, W. J., Mendro, R. L., Orsak, T. H., & Weerasinghe, D. (1998). *An application of hierarchical linear modeling to the estimation of school and teacher effect*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA, April 13-17.

Weerasinghe, D., & Orsak, T. (1998). *Can hierarchical linear modeling be used to rank schools: A simulation study with conditions under which hierarchical linear modelling is applicable*. Dallas Independent School District, Dallas, TX.

Williamson, J. B., & McNamara, T. K. (2002). *The effect of unplanned changes in marital and disability status: Interrupted trajectories and labor force participation*. A Technical Report. Chestnut Hill, MA: Center for Retirement Research at Boston College.

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An Introduction to Hierarchical Linear Modeling. *Tutorials in Quantitative Methods for Psychology*, 8, 52-69.

Wong, G. Y., & Mason, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80, 391, 513-524.

Young, D. J., Reynolds, A. J., & Walberg, H. J. (1996). Science achievement and educational productivity: A hierarchical linear model. *The Journal of Educational Research*, 89(5), 272-278.

Zeger, S., Liang, K., & Albert, P. (1988). Model for longitudinal data: A likelihood approach. *Biometrics*, 44, 1049-1060.

Zeng, L., Simonsson, M., & Poelzer, H. (2002). *Teacher certification tests: Variables that predict pass/fail status on elementary professional development examination for preservice teachers*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Zhao, L. P., & Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, 77, 642-648.

7 APPENDICES

A: SUMMARY RESULTS: EMPTY MODEL OF THE CONTINUOUS OUTCOME MODEL

```
Linear mixed model fit by REML ['lmerMod']

Formula: norm_diff3 ~ (1 | school)

   Data: df.model

REML criterion at convergence: 4406.8

Scaled residuals:

    Min      1Q  Median      3Q     Max

-2.7115 -0.7263  0.0239  0.6551  4.8193

Random effects:

 Groups   Name        Variance Std.Dev.

 school   (Intercept)  11.06    3.326

 Residual             125.40   11.198

Number of obs: 573, groups:  school, 8

Fixed effects:

            Estimate Std. Error t value

(Intercept)  -10.170      1.276  -7.968
```

B: SUMMARY RESULTS: BASELINE MODEL OF THE CONTINUOUS OUTCOME MODEL

```
print(summary(con.model1))

Linear mixed model fit by REML ['lmerMod']

Formula: norm_diff3 ~ (1 | school) + cohort + norm_diff1

   Data: df.model

REML criterion at convergence: 3747.9

Scaled residuals:

    Min      1Q  Median      3Q     Max

-2.8764 -0.6424 -0.0269  0.6261  6.4470

Random effects:

 Groups   Name        Variance Std.Dev.

 school   (Intercept)  1.565   1.251

 Residual             39.724   6.303

Number of obs: 573, groups:  school, 8

Fixed effects:

                 Estimate Std. Error t value

(Intercept)       4.10249    0.75707   5.419

cohortCohort 5-6 -2.03365    0.55475  -3.666

norm_diff1        0.82052    0.02359  34.786

Correlation of Fixed Effects:

            (Intr) chC5-6

chrtChrt5-6 -0.545

norm_diff1   0.615 -0.290
```

C: SUMMARY RESULTS: FULL MODEL OF THE CONTINUOUS OUTCOME MODEL

```
Linear mixed model fit by REML ['lmerMod']
Formula: norm_diff3 ~ (1 | school) + cohort + norm_diff1 + cpck
   Data: df.model
REML criterion at convergence: 3737.5
Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.0053 -0.6546 -0.0363  0.6188  6.4449
Random effects:
 Groups   Name        Variance Std.Dev.
 school   (Intercept)  0.7681  0.8764
 Residual             39.7060  6.3013
Number of obs: 573, groups:  school, 8
Fixed effects:
                 Estimate Std. Error t value
(Intercept)      -5.49784    4.21075  -1.306
cohortCohort 5-6 -2.00809    0.55393  -3.625
norm_diff1        0.81685    0.02351  34.746
cpck             17.92396    7.72089   2.321
Correlation of Fixed Effects:
            (Intr) chC5-6 nrm_d1
chrtChrt5-6 -0.130
norm_diff1   0.194 -0.288
cpck        -0.987  0.033 -0.086
.sig01             0.0000000  1.6181861
.sigma             5.9394854  6.6734081
(Intercept)      -13.5657625  2.5977151
cohortCohort 5-6  -3.0969618 -0.9285473
norm_diff1         0.7706035  0.8623893
cpck               3.0269945 32.7246131
```

```
> coef(summary(con.model2))
                   Estimate Std. Error        df   t value     Pr(>|t|)
(Intercept)      -5.4978392 4.21075217   7.390996 -1.305667  2.308370e-
01
cohortCohort 5-6 -2.0080867 0.55392679 568.891073 -3.625184  3.146560e-
04
norm_diff1         0.8168494 0.02350922 550.538355 34.745913 6.694195e-
141
cpck             17.9239616 7.72088540   7.031686  2.321490  5.311511e-
02
```

D: SUMMARY RESULTS: EMPTY MODEL OF THE CATEGORICAL OUTCOME MODEL

```
> print(summary(cat.model0))

Linear mixed model fit by REML ['lmerMod']

Formula: star_stan3 ~ (1 | school)

   Data: df.model

REML criterion at convergence: 2243.9

Scaled residuals:

     Min       1Q   Median       3Q      Max

-2.19021 -0.80981 -0.08118  0.74736  3.11552

Random effects:

 Groups   Name        Variance Std.Dev.

 school   (Intercept) 0.3006   0.5483

 Residual             2.8530   1.6891

Number of obs: 573, groups:  school, 8

Fixed effects:

            Estimate Std. Error t value

(Intercept)   3.7907     0.2079   18.24
```

E: SUMMARY RESULTS: BASELINE MODEL OF THE CATEGORICAL OUTCOME MODEL

```
Linear mixed model fit by REML ['lmerMod']
Formula: star_stan3 ~ (1 | school) + cohort + norm_diff1
   Data: df.model
REML criterion at convergence: 1637.9
Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.7344 -0.6582 -0.0284  0.6505  3.5497
Random effects:
 Groups   Name        Variance Std.Dev.
 school   (Intercept) 0.03545  0.1883
 Residual             0.98123  0.9906
Number of obs: 573, groups:  school, 8
Fixed effects:
                  Estimate Std. Error t value
(Intercept)       6.244773   0.117220  53.274
cohortCohort 5-6 -0.853576   0.087170  -9.792
norm_diff1        0.123925   0.003704  33.454
Correlation of Fixed Effects:
            (Intr) chC5-6
chrtChrt5-6 -0.553
norm_diff1   0.624 -0.289
```

F: SUMMARY RESULTS: FULL MODEL OF THE CATEGORICAL OUTCOME MODEL

```
Linear mixed model fit by REML ['lmerMod']
Formula: star_stan3 ~ (1 | school) + cohort + norm_diff1 + cpck
   Data: df.model
REML criterion at convergence: 1629.9
Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.7350 -0.6677 -0.0313  0.6189  3.5234
Random effects:
 Groups   Name        Variance Std.Dev.
 school   (Intercept) 0.01147  0.1071
 Residual             0.98022  0.9901
Number of obs: 573, groups:  school, 8
Fixed effects:
                  Estimate Std. Error t value
(Intercept)       4.633367   0.584920   7.921
cohortCohort 5-6 -0.851409   0.086925  -9.795
norm_diff1        0.123298   0.003678  33.525
cpck              3.014497   1.068884   2.820
Correlation of Fixed Effects:
            (Intr) chC5-6 nrm_d1
chrtChrt5-6 -0.143
norm_diff1   0.215 -0.286
cpck        -0.984  0.034 -0.093
> confint(cat.model2)
Computing profile confidence intervals ...
                      2.5 %     97.5 %
.sig01            0.0000000  0.2159663
.sigma            0.9332142  1.0485045
(Intercept)       3.5015109  5.7461094
```

```
cohortCohort 5-6 -1.0236713 -0.6831533

norm_diff1        0.1160889  0.1304315

cpck              0.9740933  5.0834441

> coef(summary(cat.model2))

                   Estimate  Std. Error        df   t value     Pr(>|t|)

(Intercept)       4.6333671 0.584920106   7.460279  7.921367  6.891739e-
05

cohortCohort 5-6 -0.8514094 0.086924847 568.986470 -9.794776  4.880819e-
21

norm_diff1        0.1232983 0.003677811 535.628070 33.524913 1.233081e-
133

cpck             3.0144972 1.068884460   7.034416  2.820227  2.563388e-
02
```

G: SUMMARY RESULTS: EMPTY MODEL OF THE BINARY OUTCOME MODEL

```
Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) [glmerMod]

 Family: binomial  ( logit )

Formula: bin_diff3 ~ 1 + (1 | school)

   Data: df.model

Control: my.control

     AIC      BIC   logLik deviance df.resid

   501.3    510.0   -248.6    497.3      571

Scaled residuals:

    Min      1Q  Median      3Q     Max

-0.5312 -0.4759 -0.3766 -0.3395  2.9454

Random effects:

 Groups Name        Variance Std.Dev.

 school (Intercept) 0.1635   0.4044

Number of obs: 573, groups:  school, 8

Fixed effects:

            Estimate Std. Error z value Pr(>|z|)

(Intercept)  -1.7161     0.1893  -9.064   <2e-16 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

H: SUMMARY RESULTS: BASELINE MODEL OF THE BINARY OUTCOME MODEL

```
Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: bin_diff3 ~ 1 + (1 | school) + cohort + norm_diff1
   Data: df.model
Control: my.control
     AIC      BIC   logLik deviance df.resid
   296.4    313.8   -144.2    288.4      569
Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.4283 -0.2940 -0.1129 -0.0323  4.2550
Random effects:
 Groups Name        Variance Std.Dev.
 school (Intercept) 0.05746  0.2397
Number of obs: 573, groups:  school, 8
Fixed effects:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.92542    0.35706   2.592  0.00955 **
cohortCohort 5-6 -0.68003    0.33863  -2.008  0.04463 *
norm_diff1        0.23597    0.02583   9.135  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Correlation of Fixed Effects:
            (Intr) chC5-6
chrtChrt5-6 -0.769
norm_diff1   0.661 -0.366
```

I: SUMMARY RESULTS: FULL MODEL OF THE BINARY OUTCOME MODEL

```
Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: bin_diff3 ~ 1 + (1 | school) + cohort + norm_diff1 + cpck
   Data: df.model
Control: my.control
     AIC      BIC   logLik deviance df.resid
   295.3    317.1   -142.7    285.3      568
Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.9482 -0.2935 -0.1122 -0.0319  4.4288
Random effects:
 Groups Name        Variance Std.Dev.
 school (Intercept) 0        0
Number of obs: 573, groups:  school, 8
Fixed effects:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.95446    1.58524  -1.233   0.2176
cohortCohort 5-6 -0.59913    0.33980  -1.763   0.0779 .
norm_diff1        0.23227    0.02512   9.246   <2e-16 ***
cpck              5.24023    2.82999   1.852   0.0641 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Correlation of Fixed Effects:
            (Intr) chC5-6 nrm_d1
chrtChrt5-6 -0.291
norm_diff1   0.110 -0.357
cpck        -0.976  0.124  0.038
convergence code: 0
```

J: SUMMARY RESULTS: EMPTY MODEL OF THE 2-LEVEL MODEL

```
Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: pass ~ 1 + (1 | provider_id)
   Data: df.model
Control: my.control

     AIC      BIC   logLik deviance df.resid
 54022.5  54039.8 -27009.3  54018.5    42171

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.7694 -1.1260  0.6315  0.8011  1.6111

Random effects:
 Groups      Name        Variance Std.Dev.
 provider_id (Intercept) 0.2949   0.543
Number of obs: 42173, groups:  provider_id, 34

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.31244    0.09523   3.281  0.00103 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

K: SUMMARY RESULTS: FULL MODEL OF THE 2-LEVEL MODEL

```
Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: pass ~ (1 | provider_id) + ethnicity + subject + nat_pass +
ld_coverage +
    intervention + decile + time + ld_coverage:intervention
   Data: df.model
Control: my.control

     AIC      BIC   logLik deviance df.resid
 52938.7  53051.2 -26456.4  52912.7    42160

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.2146 -1.0452  0.5642  0.7522  2.1585

Random effects:
 Groups      Name        Variance Std.Dev.
 provider_id (Intercept) 0.09079  0.3013
Number of obs: 42173, groups:  provider_id, 34

Fixed effects:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.27902    0.21356  -5.989 2.11e-09 ***
ethnicityOther           0.58630    0.03091  18.968  < 2e-16 ***
ethnicityPasifika       -0.12200    0.03769  -3.237 0.001209 **
subjectMathematics      -0.46943    0.04534 -10.354  < 2e-16 ***
subjectScience          -0.58521    0.03938 -14.861  < 2e-16 ***
nat_pass                 3.42623    0.30946  11.072  < 2e-16 ***
ld_coverage             -0.13323    0.05555  -2.398 0.016470 *
```

```
intervention             -0.25907    0.04718  -5.491 4.01e-08 ***
decileLowest             -0.69044    0.13589  -5.081 3.76e-07 ***
decileMedium             -0.21127    0.13297  -1.589 0.112100
time                      0.05371    0.01516   3.544 0.000394 ***
ld_coverage:intervention  0.24822    0.06657   3.729 0.000192 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) ethncO ethncP sbjctM sbjctS nt_pss ld_cvr intrvn
dclLws dclMdm time
ethnctyOthr -0.124
ethnctyPsfk -0.069  0.611
sbjctMthmtc  0.518 -0.037  0.011
subjectScnc  0.546 -0.062  0.018  0.780
nat_pass    -0.862  0.017 -0.013 -0.725 -0.720
ld_coverage -0.362 -0.003  0.008  0.230  0.083  0.184
interventin -0.390 -0.012 -0.004 -0.148 -0.204  0.314  0.493
decileLowst -0.321  0.055 -0.041  0.000 -0.010  0.010  0.037  0.026
decileMedim -0.296   0.005 -0.025 -0.002 -0.002 -0.010 -0.017 -0.011
0.487
time         0.547  0.030  0.003  0.402  0.420 -0.584 -0.070 -0.506 -
0.001  0.019
ld_cvrg:ntr  0.248  0.002  0.005  0.105  0.155 -0.199 -0.471 -0.807 -
0.031  0.005  0.075
convergence code: 0
Model failed to converge with max|grad| = 0.00298081 (tol = 0.001,
component 1)

> anova(lmfit20, lmfit21)
Data: df.model
Models:
lmfit20: pass ~ 1 + (1 | provider_id)
```

```
lmfit21: pass ~ (1 | provider_id) + ethnicity + subject + nat_pass +
ld_coverage +
lmfit21:     intervention + decile + time + ld_coverage:intervention
        Df   AIC   BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
lmfit20  2 54023 54040 -27009    54019
lmfit21 13 52939 53051 -26456    52913 1105.8     11  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

L: SUMMARY RESULTS: EMPTY MODEL OF THE 3-LEVEL-D MODEL

```
Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: pass ~ 1 + (1 | provider_id) + (1 | provider_id:subject)
   Data: df.model
Control: my.control

     AIC      BIC   logLik deviance df.resid
 53647.4  53673.4 -26820.7  53641.4    42170

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.1569 -1.1249  0.5894  0.7658  2.1916

Random effects:
 Groups              Name        Variance Std.Dev.
 provider_id:subject (Intercept) 0.1155   0.3399
 provider_id         (Intercept) 0.2859   0.5347
Number of obs: 42173, groups:  provider_id:subject, 100; provider_id,
34

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3123     0.1002   3.116  0.00183 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> anova(lmfit20, lmfit3a, lmfit40)
```

```
Data: df.model

Models:

lmfit20: pass ~ 1 + (1 | provider_id)

lmfit3a: pass ~ 1 + (1 | provider_id) + (1 | provider_id:subject)

lmfit40: pass ~ 1 + (1 | provider_id) + (1 | provider_id:subject) + (1
|

lmfit40:     nsn)

        Df   AIC   BIC logLik deviance  Chisq Chi Df Pr(>Chisq)

lmfit20  2 54023 54040 -27009    54019

lmfit3a  3 53647 53673 -26821    53641  377.1      1  < 2.2e-16 ***

lmfit40  4 51468 51503 -25730    51460 2181.1      1  < 2.2e-16 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

M: SUMMARY RESULTS: FULL MODEL OF THE 3-LEVEL-D MODEL

```
Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: pass ~ (1 | provider_id) + (1 | provider_id:subject) + ethnicity
+
    nat_pass + subject + ld_coverage + intervention + decile +
    time + ld_coverage:intervention
   Data: df.model
Control: my.control

     AIC      BIC   logLik deviance df.resid
 52623.6  52744.7 -26297.8  52595.6    42159

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.4540 -1.0377  0.5606  0.7461  2.6821

Random effects:
 Groups              Name        Variance Std.Dev.
 provider_id:subject (Intercept) 0.10053  0.3171
 provider_id         (Intercept) 0.06838  0.2615
Number of obs: 42173, groups:  provider_id:subject, 100; provider_id,
34

Fixed effects:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            -1.18443    0.22552  -5.252 1.50e-07 ***
ethnicityOther          0.58502    0.03111  18.808  < 2e-16 ***
ethnicityPasifika      -0.12090    0.03796  -3.185  0.00145 **
nat_pass                3.39130    0.30820  11.004  < 2e-16 ***
```

```
subjectMathematics       -0.50547    0.09391  -5.383 7.35e-08 ***
subjectScience           -0.62566    0.09160  -6.830 8.48e-12 ***
ld_coverage              -0.15097    0.07294  -2.070  0.03848 *
intervention             -0.31070    0.05200  -5.975 2.30e-09 ***
decileLowest             -0.73234    0.14393  -5.088 3.62e-07 ***
decileMedium             -0.23135    0.14051  -1.646  0.09966 .
time                      0.06448    0.01523   4.234 2.30e-05 ***
ld_coverage:intervention  0.32121    0.07484   4.292 1.77e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr)  ethncO  ethncP  nt_pss  sbjctM  sbjctS  ld_cvr  intrvn
dclLws dclMdm time
ethnctyOthr -0.114
ethnctyPsfk -0.062  0.610
nat_pass    -0.827  0.013 -0.015
sbjctMthmtc  0.047 -0.015  0.005 -0.328
subjectScnc  0.058 -0.026  0.004 -0.305  0.545
ld_coverage -0.409 -0.004 -0.003  0.200  0.145  0.035
interventin -0.417 -0.015 -0.008  0.327 -0.035 -0.071  0.536
decileLowst -0.324  0.052 -0.038  0.010  0.003 -0.004  0.040  0.029
decileMedim -0.292   0.004 -0.023 -0.012 -0.004 -0.010 -0.024 -0.014
0.486
time         0.512  0.032  0.003 -0.575  0.182  0.176 -0.067 -0.478 -
0.002  0.020
ld_cvrg:ntr  0.290  0.007  0.008 -0.225  0.023  0.052 -0.495 -0.837 -
0.032  0.007  0.084
convergence code: 0
Model failed to converge with max|grad| = 0.00451038 (tol = 0.001,
component 1)

Warning message:
```

```
Unknown or uninitialised column: 'timepoint'.
> anova(lmfit3a, lmfit3f)
Data: df.model
Models:
lmfit3a: pass ~ 1 + (1 | provider_id) + (1 | provider_id:subject)
lmfit3f: pass ~ (1 | provider_id) + (1 | provider_id:subject) + ethnicity
+
lmfit3f:     nat_pass + subject + ld_coverage + intervention + decile +
lmfit3f:     time + ld_coverage:intervention
        Df   AIC   BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
lmfit3a  3 53647 53673 -26821    53641
lmfit3f 14 52624 52745 -26298    52596 1045.9     11  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> anova(lmfit21, lmfit3f, lmfit41)
Data: df.model
Models:
lmfit21: pass ~ (1 | provider_id) + ethnicity + subject + nat_pass +
ld_coverage +
lmfit21:     intervention + decile + time + ld_coverage:intervention
lmfit3f: pass ~ (1 | provider_id) + (1 | provider_id:subject) + ethnicity
+
lmfit3f:     nat_pass + subject + ld_coverage + intervention + decile +
lmfit3f:     time + ld_coverage:intervention
lmfit41: pass ~ (1 | provider_id) + (1 | provider_id:subject) + (1 |
nsn) +
lmfit41:     ethnicity + nat_pass + subject + ld_coverage + intervention
+
lmfit41:     decile + time + ld_coverage:intervention
        Df   AIC   BIC logLik deviance   Chisq Chi Df Pr(>Chisq)
lmfit21 13 52939 53051 -26456    52913
```

```
lmfit3f 14 52624 52745 -26298    52596  317.17      1  < 2.2e-16 ***

lmfit41 15 50621 50751 -25296    50591 2004.63      1  < 2.2e-16 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

N: SUMMARY RESULTS: EMPTY MODEL OF THE 3-LEVEL-R MODEL

```
Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: pass ~ 1 + (1 | provider_id) + (1 | nsn)
   Data: df.model
Control: my.control

     AIC      BIC   logLik deviance df.resid
 51998.4  52024.3 -25996.2  51992.4    42170

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.1109 -0.7803  0.4239  0.6221  1.8605

Random effects:
 Groups      Name        Variance Std.Dev.
 nsn         (Intercept) 1.3850   1.1769
 provider_id (Intercept) 0.4401   0.6634
Number of obs: 42173, groups:  nsn, 15090; provider_id, 34

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.2627     0.1177   2.233   0.0256 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

O: SUMMARY RESULTS: FULL MODEL OF THE 3-LEVEL-R MODEL

```
Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) [glmerMod]

 Family: binomial  ( logit )

Formula: pass ~ (1 | provider_id) + (1 | nsn) + ethnicity + subject +

    nat_pass + ld_coverage + intervention + decile + time +
ld_coverage:intervention

   Data: df.model

Control: my.control


     AIC      BIC   logLik deviance df.resid

 50928.5  51049.6 -25450.3  50900.5    42159


Scaled residuals:

    Min      1Q  Median      3Q     Max

-2.8759 -0.7338  0.3985  0.6097  2.1550


Random effects:

 Groups      Name        Variance Std.Dev.

 nsn         (Intercept) 1.4590   1.2079

 provider_id (Intercept) 0.1134   0.3367

Number of obs: 42173, groups:  nsn, 15090; provider_id, 34


Fixed effects:

                        Estimate Std. Error z value Pr(>|z|)

(Intercept)             -1.85495    0.25171  -7.369 1.71e-13 ***

ethnicityOther           0.77694    0.04735  16.407  < 2e-16 ***

ethnicityPasifika       -0.16784    0.05716  -2.936  0.00332 **

subjectMathematics      -0.67010    0.05352 -12.520  < 2e-16 ***

subjectScience          -0.95806    0.04755 -20.149  < 2e-16 ***

nat_pass                 4.65804    0.36280  12.839  < 2e-16 ***
```

```
ld_coverage              -0.21937    0.06852  -3.202  0.00137 **
intervention             -0.34543    0.06275  -5.505 3.69e-08 ***
decileLowest             -0.88054    0.15721  -5.601 2.13e-08 ***
decileMedium             -0.29010    0.15183  -1.911  0.05604 .
time                      0.03977    0.02146   1.854  0.06380 .
ld_coverage:intervention  0.37025    0.08182   4.525 6.04e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Correlation of Fixed Effects:
            (Intr) ethncO ethncP sbjctM sbjctS nt_pss ld_cvr intrvn
dclLws dclMdm time
ethnctyOthr -0.162
ethnctyPsfk -0.091  0.597
sbjctMthmtc  0.505 -0.044  0.011
subjectScnc  0.534 -0.073  0.019  0.782
nat_pass    -0.860  0.023 -0.013 -0.719 -0.713
ld_coverage -0.371 -0.007  0.005  0.248  0.103  0.183
interventin -0.381 -0.013 -0.003 -0.127 -0.173  0.285  0.452
decileLowst -0.311  0.071 -0.058  0.004 -0.004  0.009  0.041  0.024
decileMedim -0.285   0.008 -0.033 -0.003 -0.001 -0.009 -0.017 -0.010
0.481
time          0.481  0.028  0.000  0.328  0.340 -0.485 -0.059 -0.556
0.000  0.022
ld_cvrg:ntr  0.256  0.002  0.003  0.108  0.151 -0.204 -0.470 -0.743 -
0.034  0.004  0.061
> anova(lmfit30, lmfit31)
Data: df.model
Models:
lmfit30: pass ~ 1 + (1 | provider_id) + (1 | nsn)
lmfit31: pass ~ (1 | provider_id) + (1 | nsn) + ethnicity + subject +
lmfit31:      nat_pass + ld_coverage + intervention + decile + time +
ld_coverage:intervention
```

```
        Df   AIC   BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
lmfit30  3 51998 52024 -25996    51992
lmfit31 14 50929 51050 -25450    50901 1091.9     11  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

P: SUMMARY RESULTS: EMPTY MODEL OF THE 4-LEVEL MODEL

```
Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) [glmerMod]

 Family: binomial  ( logit )

Formula: pass ~ 1 + (1 | provider_id) + (1 | provider_id:subject) + (1
|      nsn)

   Data: df.model

Control: my.control


     AIC      BIC   logLik deviance df.resid

 51468.3  51502.9 -25730.2  51460.3    42169


Scaled residuals:

    Min      1Q  Median      3Q     Max

-2.5500 -0.7387  0.4117  0.6100  2.3324


Random effects:

 Groups              Name        Variance Std.Dev.

 nsn                 (Intercept) 1.5462   1.2435

 provider_id:subject (Intercept) 0.2231   0.4723

 provider_id         (Intercept) 0.4540   0.6738

Number of obs: 42173, groups:  nsn, 15090; provider_id:subject, 100;
provider_id, 34


Fixed effects:

            Estimate Std. Error z value Pr(>|z|)

(Intercept)   0.2321     0.1290     1.8   0.0719 .

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> anova(lmfit20, lmfit30, lmfit40)

Data: df.model

Models:

lmfit20: pass ~ 1 + (1 | provider_id)

lmfit30: pass ~ 1 + (1 | provider_id) + (1 | nsn)

lmfit40: pass ~ 1 + (1 | provider_id) + (1 | provider_id:subject) + (1
|

lmfit40:     nsn)

        Df   AIC   BIC logLik deviance   Chisq Chi Df Pr(>Chisq)

lmfit20  2 54023 54040 -27009    54019

lmfit30  3 51998 52024 -25996    51992 2026.15      1  < 2.2e-16 ***

lmfit40  4 51468 51503 -25730    51460  532.09      1  < 2.2e-16 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Q: SUMMARY RESULTS: FULL MODEL OF THE 4-LEVEL MODEL

```
Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: pass ~ (1 | provider_id) + (1 | provider_id:subject) + (1 |
nsn) +
    ethnicity + nat_pass + subject + ld_coverage + intervention +
    decile + time + ld_coverage:intervention
   Data: df.model
Control: my.control

     AIC      BIC   logLik deviance df.resid
 50620.9  50750.7 -25295.5  50590.9    42158

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.9514 -0.7234  0.3911  0.6041  2.6559

Random effects:
 Groups              Name        Variance Std.Dev.
 nsn                 (Intercept) 1.47836  1.2159
 provider_id:subject (Intercept) 0.14065  0.3750
 provider_id         (Intercept) 0.09288  0.3048
Number of obs: 42173, groups:  nsn, 15090; provider_id:subject, 100;
provider_id, 34

Fixed effects:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.69756    0.27667  -6.136 8.47e-10 ***
ethnicityOther           0.77872    0.04769  16.327  < 2e-16 ***
ethnicityPasifika       -0.16974    0.05766  -2.944  0.00324 **
```

```
nat_pass                   4.54816    0.37435  12.150  < 2e-16 ***
subjectMathematics        -0.71404    0.11171  -6.392 1.64e-10 ***
subjectScience            -1.00213    0.10939  -9.161  < 2e-16 ***
ld_coverage               -0.24606    0.09239  -2.663  0.00774 **
intervention              -0.42078    0.06901  -6.098 1.08e-09 ***
decileLowest              -0.94658    0.17314  -5.467 4.58e-08 ***
decileMedium              -0.31391    0.16759  -1.873  0.06105 .
time                       0.05795    0.02178   2.661  0.00779 **
ld_coverage:intervention   0.46808    0.09297   5.035 4.79e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Correlation of Fixed Effects:
            (Intr)  ethncO  ethncP  nt_pss  sbjctM  sbjctS  ld_cvr  intrvn
dclLws dclMdm time
ethnctyOthr -0.148
ethnctyPsfk -0.081  0.596
nat_pass    -0.827  0.022 -0.013
sbjctMthmtc  0.048 -0.020  0.004 -0.329
subjectScnc  0.067 -0.032  0.005 -0.313  0.549
ld_coverage -0.427 -0.005 -0.004  0.214  0.154  0.038
interventin -0.415 -0.015 -0.006  0.315 -0.033 -0.067  0.496
decileLowst -0.318  0.066 -0.051  0.013  0.005 -0.002  0.041  0.029
decileMedim -0.283   0.008 -0.029 -0.011 -0.005 -0.008 -0.025 -0.013
0.482
time         0.459  0.029  0.000 -0.489  0.155  0.152 -0.065 -0.531 -
0.003  0.020
ld_cvrg:ntr  0.303  0.006  0.006 -0.240  0.028  0.056 -0.486 -0.785 -
0.035  0.006  0.080
> anova(lmfit40, lmfit41)
Data: df.model
Models:
```

```
lmfit40: pass ~ 1 + (1 | provider_id) + (1 | provider_id:subject) + (1
|
lmfit40:     nsn)
lmfit41: pass ~ (1 | provider_id) + (1 | provider_id:subject) + (1 |
nsn) +
lmfit41:     ethnicity + nat_pass + subject + ld_coverage + intervention
+
lmfit41:     decile + time + ld_coverage:intervention
        Df   AIC   BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
lmfit40  4 51468 51503 -25730    51460
lmfit41 15 50621 50751 -25296    50591 869.36     11  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> anova(lmfit21, lmfit31, lmfit41)
Data: df.model
Models:
lmfit21: pass ~ (1 | provider_id) + ethnicity + subject + nat_pass +
ld_coverage +
lmfit21:     intervention + decile + time + ld_coverage:intervention
lmfit31: pass ~ (1 | provider_id) + (1 | nsn) + ethnicity + subject +
lmfit31:      nat_pass + ld_coverage + intervention + decile + time +
ld_coverage:intervention
lmfit41: pass ~ (1 | provider_id) + (1 | provider_id:subject) + (1 |
nsn) +
lmfit41:     ethnicity + nat_pass + subject + ld_coverage + intervention
+
lmfit41:     decile + time + ld_coverage:intervention
        Df   AIC   BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
lmfit21 13 52939 53051 -26456    52913
lmfit31 14 50929 51050 -25450    50901 2012.2      1  < 2.2e-16 ***
lmfit41 15 50621 50751 -25296    50591  309.6      1  < 2.2e-16 ***
```