Using PAT: Reading Comprehension data to understand student progress in reading

Melanie Berg and Elliot Lawes

NZCER
Rangahau Mātauranga o Aotearoa

New Zealand Council for Educational Research
PO Box 3237
Wellington
New Zealand
www.nzcer.org.nz

ISBN 978-0-947509-49-1

Contents

Tables

Figures

Summary

To date, large-scale national monitoring of student reading in Years 4 to 10 in New Zealand has been limited to describing achievement. We use multilevel modelling to summarise student *progress* using data captured between 2008 and 2015 by the Progressive Achievement Tests in Reading Comprehension.

Broadly speaking, average student progress in reading occurs at a similar rate regardless of student gender, or ethnic group and school decile.

More specifically:

- We confirm the average differences in reading achievement according to gender, ethnic group and school decile at Years 4, 5 and 8 that have been found in other studies.[1]
- We find that while there are some statistically significant differences in rates of progress in reading according to gender, ethnic group and school decile, these differences are small.
- In particular, we find that, overall, boys and students in decile 1–2 schools tend to have lower achievement in Year 4 but tend to make slightly faster progress than other students from Year 4 to Year 10.

This report is presented in two parts. Part A is an account of our research that does not include technical detail. Part B provides the technical and methodological detail underpinning the research. Part A can be read independently of Part B.

[1] The Progress in International Reading Literacy Study (PIRLS) (Chamberlain, 2014) and the National Monitoring Study of Student Achievement (NMSSA) (Educational Assessment Research Unit & New Zealand Council for Educational Research, 2016).

Part A: Progress in reading

This part of the report is an account of our research that does not include technical detail.

1. Large-scale monitoring of reading achievement in New Zealand

Over the past quarter of a century a number of relatively large-scale exercises that monitor reading achievement in English medium nationally have been carried out in New Zealand. In the adult sector there is the International Adult Literacy study (Walker, Udy, & Pole, 1996), the Adult Literacy and Life-skills study (Satherley & Lawes, 2007) and the Programme for the International Assessment of Adult Competencies (Ministry of Education, 2016). In the secondary school sector there is the Programme for International Student Assessment (PISA) (Telford & May, 2010). In the primary school sector there is the Progress in International Reading Literacy Study (PIRLS) (Chamberlain, 2014), the National Education Monitoring Project (NEMP) (Gilmore & Smith, 2011) and the National Monitoring Study of Student Achievement (NMSSA) (Educational Assessment Research Unit & New Zealand Council for Educational Research, 2016). All of these studies are funded by the New Zealand Government, often in partnership with an international agency such as the Organisation for Economic Co-operation and Development (OECD) in the case of PISA.

The findings of these exercises have been used to inform a number of policy developments and debates. For example, the development of the National Standards in Reading (Ministry of Education, 2013b) drew heavily on the Literacy Learning Progressions (Ministry of Education, 2010c) which, in turn, used PISA, PIRLS and NEMP to frame what reading achievement might mean and to gauge student reading achievement in New Zealand (Ministry of Education, 2010a, 2010b).

Another example of the impact of the large-scale monitoring of reading achievement in New Zealand is the use of findings from the PIRLS study in the debate around Reading Recovery. Reading Recovery is a programme—developed by Dame Marie Clay in the late 1970s and first implemented in the mid-1980s—that seeks to lift the reading and writing achievement of students who have made less-than-expected progress after one year of schooling (Research Division, Ministry of Education, 2014). Both internationally and in New Zealand, research about Reading Recovery is a contested area where scholars and educators debate, among other issues, the merits of using whole-language versus phonics-based approaches to teaching reading (for example, Soler & Openshaw, 2006). In New Zealand, Tunmer, Chapman, Greaney, Prochnow and Arrow (2013) use evidence from PIRLS in their argument that Reading Recovery—as part of New Zealand's Literacy Strategy—is not working as an educational intervention nationally. In particular, Tunmer et al. contrast the extent of New Zealand's implementation of Reading Recovery with the lack of change in the distribution of reading achievement of Year 5 students since 2001 as measured by PIRLS (2013). Whether or not readers are convinced by the argument of Tunmer et al., the fact remains that the PIRLS study provides key information about reading achievement in New Zealand that is used to inform debate nationally.

A common feature of all of the large-scale exercises that monitor reading achievement in English nationally in New Zealand is that they explicitly describe reading achievement in various educational sectors at a given time. However, because many of these exercises are repeated every so many years (e.g. PIRLS), or collect data from students from several year levels (e.g. NMSSA), they also report on changes in reading achievement over time or between year levels. This inevitably leads to questions about student *progress* in reading that policy makers attempt to address. For example, the National Standards policy—informed by findings from the PISA, PIRLS and NEMP studies of reading achievement—has a strong focus on progress (Ministry of Education, 2013a). In doing this, policy makers are making unverified inferences about progress based on cross-sectional studies of reading *achievement*.

Figure 1 allows us to explore the nature of these inferences more fully. The figure shows the average reading comprehension scores of two groups of students—Group 1 and Group 2—where Group 1 is assessed at Time 1 and Group 2 is assessed at Time 2. The figure also shows the difference between the average scores for the two groups and the elapsed time between Time 1 and Time 2.

To understand Figure 1, it is useful to consider a population of students about whom policy makers might like to understand progress in reading comprehension. Because progress is a concept that depends upon time, this population shouldn't be characterised by characteristics that change with time, such as year level or calendar year. An example of such a population is Pasifika girls.

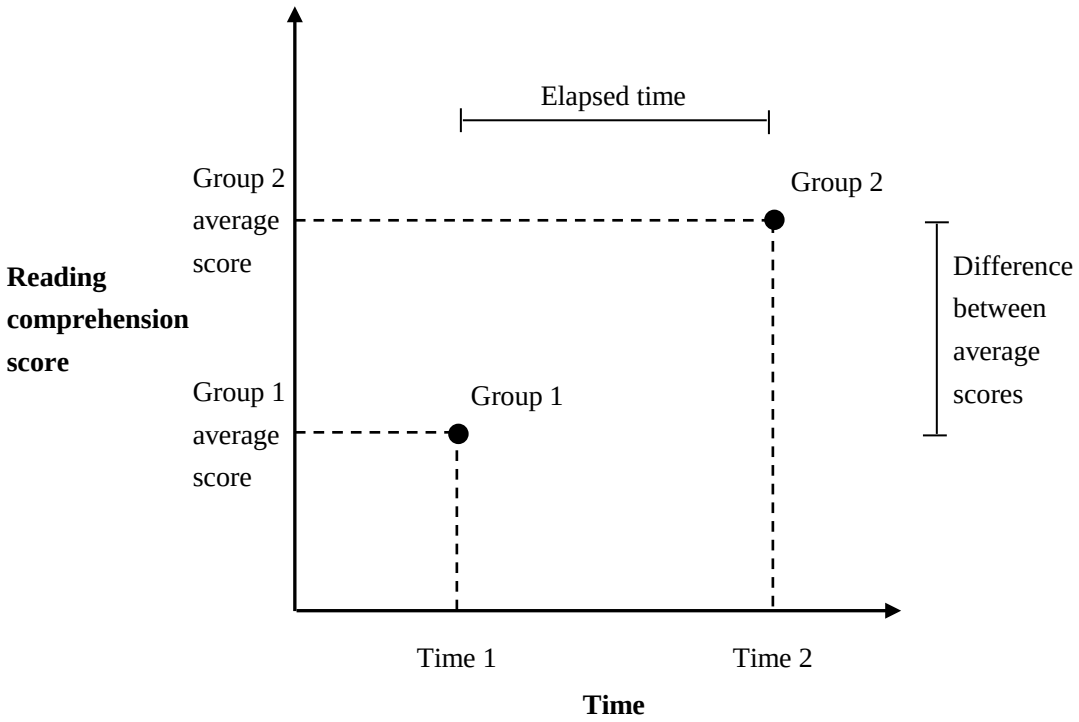
When using an achievement study for students in a specified year level that is repeated every few years (e.g. PIRLS) to make inferences about the average progress of our population, it is typical practice to assume that the progress made by our population over the elapsed time is the difference in average scores shown in Figure 1. This inference ignores the fact that Group 1 and Group 2 are distinct groups of students and confuses the cohorts of our population that Group 1 and Group 2 represent.

When using an achievement study for students in several year levels (e.g. NMSSA) to make inferences about the average progress of our population, Time 1 and Time 2 in Figure 1 are actually the same, but Group 2 represents students in our population who are older than the students in Group 1 by the elapsed time. This inference also ignores the fact that Group 1 and Group 2 are distinct groups of students and assumes that achievement will be stable for the duration of the elapsed time after Time 1.

In contrast to the two types of inference just described, when using a longitudinal study to understand progress, Group 1 and Group 2 are the same students assessed at different times. Therefore the average progress of this group of students over the elapsed time *is* the difference in average scores shown in Figure 1.

There are few recent examples of large-scale reading research or monitoring in New Zealand that actually seek to understand progress in reading. Lai, McNaughton, Amituanai-Toloa, Turner and Hsiao (2009) describe the results of a schooling improvement study aimed at accelerating achievement in reading. This 3-year study measured the reading progress of the students in seven schools in low socioeconomic communities and was never intended to describe progress in reading nationally. Of course, if large-scale exercises that monitor reading *achievement* nationally are expensive, then exercises of a similar scale that monitor reading *progress* nationally are prohibitively so, as they require all of the work demanded by an achievement monitoring study replicated several times with the additional issue of tracking students' changing environments over time.

Figure 1 **Understanding the nature of inferences about student progress made using student achievement data**



But despite this lack of evidence about progress, policy demand remains and currently policy makers in New Zealand are in the unenviable position of having to make inferences about progress in reading based on information about achievement in reading. Lawes (2016) shows how the use of achievement information in the development of National Standards policy on progress could result in distributions of student progress that challenge our intuition.

The current paper addresses the lack of large-scale quantitative research that describes progress in reading nationally in New Zealand.

2. Aim and research questions

The overall aim of the research informing this paper is to understand progress in reading for students in Years 4 to 10. We use large-scale data captured in the administration of the PAT: Reading Comprehension assessment (Darr, McDowall, Ferral, Twist, & Watson, 2008). We have approached this research quantitatively and have framed two specific research questions in that context—one focused on achievement and one on progress.

Our answer to the achievement question was gifted to us in an intermediate stage of the methodology we used to address the progress question. While relevant to the policy debate around reading in New Zealand, and certainly worth reporting, we consider the achievement question of lesser importance than the progress question.

2.1. How does the variation in student reading achievement within a school compare with that between schools?

Our first research question focuses on variation in the measure of reading achievement. The seemingly technical issue of variation in the measure of reading achievement has been ushered into the realm of policy makers and practitioners by the PISA study. For example, OECD (2010) reports that New Zealand has a wide spread of reading achievement when compared with other countries. Furthermore, the variance in reading achievement due to differences in school is much smaller than the variance due to difference in students (OECD, 2010, Figure II.5.1 and Table II.5.1). Another way of saying this is that there is greater variation in student achievement within schools than there is between schools. These facts are often interpreted as a reflection of the need for each school to cater to diverse learners (Nusche, Laveault, MacBeath, & Santiago, 2012; Robinson, Hohepa, & Lloyd, 2015, p. 57).

In the absence of any analogous New Zealand-based research findings in the primary sector, we take our guidance from PISA and hypothesise that, after accounting for residual variance, the variance in reading achievement (as measured by the PAT: Reading Comprehension assessment) due to differences in school is much smaller than the variance due to difference in students.

2.2. Does student progress in reading occur at substantially different rates according to gender, ethnic group and school decile?

Our second, and main, research question probes the extent to which the differences in average reading achievement for different subpopulations increase with time. Framed slightly differently, this question investigates the extent to which Matthew effects (Stanovich, 1986, p. 381) are associated with the demographic backgrounds of students. Matthew effects are the phenomenon where those who initially possess a comparatively large amount of an attribute are able to further acquire the attribute more readily than those who initially possess a comparatively small amount. In this case, the attribute is reading comprehension.

Hypothesising around this question is compromised by the lack of large-scale data focused on the reading progress of New Zealand's primary school students. However, we make do using results from some of New

Zealand's large-scale reading achievement monitoring projects. If we look first to patterns of change in the PIRLS study, which collects and reports on data for Year 5 students every 5 years, we would hypothesise that there were at most small differences in rates of progress with Māori boys and New Zealand European boys perhaps making slightly faster progress than other groups (Chamberlain, 2014, p. 10).

If we look to the NMSSA study, which collects and reports on data from students in Years 4 and 8, we can use the differences in achievement between students from different year levels in the study as a proxy for progress between those year levels. In this case we would hypothesise that Asian students had made less progress than other students (Educational Assessment Research Unit & New Zealand Council for Educational Research, 2016, p. 4).

3. Methodology

This section summarises the technical content of our research. Full technical details are in Part B of this report.

3.1. A measure of reading comprehension

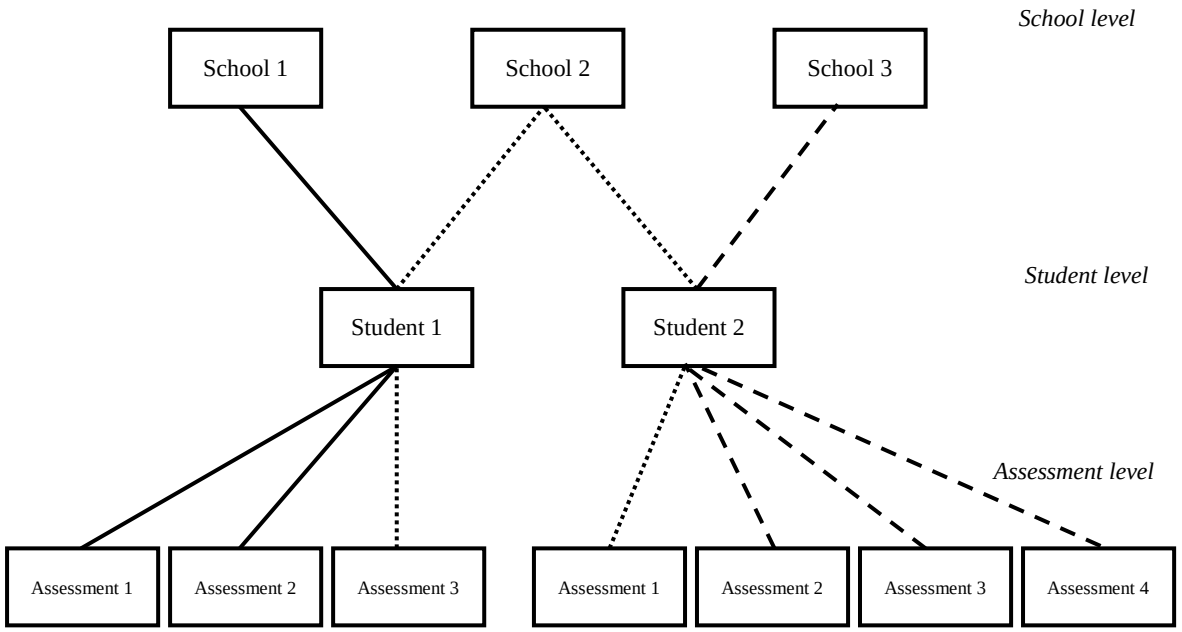
The PAT: Reading Comprehension assessment is a sequence of standardised, low-stakes tests developed to assess the reading comprehension of students in Years 4 to 10 in New Zealand schools (Darr et al., 2008). Schools choose to use the PAT: Reading Comprehension assessment and choose when and how they use it.

The PAT: Reading Comprehension tests are all linked to an equal interval scale known as the PAT: Reading Comprehension scale. The fact that the scale is an equal interval scale means that progress of one unit up the scale indicates the same amount of progress in reading no matter where on the scale the progress occurs. Units on the PAT: Reading Comprehension scale are known as *patc* units.

3.2. Data

The PAT: Reading Comprehension data has a 'multilevel structure'. Briefly, this means that information associated with each assessment a student attempts (the 'assessment level') is linked to information associated with the student (the 'student level') which is in turn linked to information about the school where the student attempted the assessment (the 'school level'). Moreover, the multilevel structure of the PAT: Reading Comprehension data is 'cross-classified'. That is, each assessment instance is associated with one student and one school, but individual students can be associated with different schools. Assessment instances are linked to a school through the student who completed it. An example of this cross-classified multilevel structure is displayed in Figure 2. In the figure, Student 1 completes their first and second assessments at School 1 and their third assessment at School 2. Student 2 completes their first assessment at School 2 and their second, third and fourth assessments at School 3.

Figure 2 **An example of the cross-classified multilevel structure of the PAT: Reading Comprehension data**



Because the PAT: Reading Comprehension data is an administrative dataset, it contains only limited information at each level. For example, at the assessment level it contains PAT: Reading Comprehension score and time of assessment, at the student level it contains demographic information and at the school level it contains administrative information.

The PAT: Reading Comprehension assessment is administered in a wide range of schools. Consequently, while there are strong protocols that maintain the quality of the data from this assessment, there are opportunities for data entry error. For example, a student's name might be recorded with different spellings on different assessment instances, or their gender might be recorded in some assessment instances, but not others. Therefore, the data was cleaned extensively before any analysis.

Cleaning of the PAT: Reading Comprehension data involved several stages. The first stage of data cleaning involved fuzzy matching to link student records longitudinally. Much of the data was already longitudinally linked, but the linking variable (National Student Number—see Ministry of Education, 2015b) was used less extensively in the earlier years of data collection. Further stages of data cleaning focused on consistency and grouping of data at the student and assessment levels. The final stage of data cleaning involved removing assessment records for students who had been assessed fewer than four times. This reduced the size of the data substantially (from 864,632 to 352,473 assessment records), but was necessary to be able to statistically model reading progress in a valid way.

The resulting data consisted of 352,473 assessment records of 70,505 students at 716 schools, where each student had been assessed four or more times between 2008 and 2015 when they were in Years 4 through 10. Table 1 in Section 7.1 summarises this data.

3.3. Analysis

Following an intial exploration of the data, we use two multilevel linear models to summarise the features of the PAT: Reading Comprehension data and address our research questions. Multilevel linear models are applied to data with a multilevel structure to determine the value of a dependent variable (such as reading achievement), based on the values of the independent variables (such as school characteristics and student characteristics).

We use one of our models (the 'null model') to address the research question "How does the variation in student reading achievement within a school compare with that between schools?" We use the other, more complex model (the 'main model') to address the research question "Does student progress in reading occur at substantially different rates according to gender, ethnic group and school decile?" Our multilevel models account for the way that PAT: Reading Comprehension scale score (an assessment level variable) varies by student and school.

We used the software environment R for all of our statistical analysis and, in particular, for multilevel modelling we used the R package 'lme4' developed by Bates, Maechler, Bolker and Walker (2015) and described in Finch, Bolin and Kelley (2014).

3.4. Results

The most important outputs from our modelling process are the parameter estimates for each model. These are numerical quantities—presented in Table 2 and Table 4 in Section 9—that describe:

- the association between each variable in our model and PAT: Reading Comprehension
- the amount of variation in the data at the student and school data levels
- the amount of variation in the data that the model does not explain.

To address the question "How does the variation in student reading achievement within schools compare with that between schools?" we used the parameter estimates from our null model (Table 2) to compare the amount of variation in the data at the student and school data levels after accounting for the amount of variation in the data that the model does not explain.

To address the question "Does student progress in reading occur at substantially different rates according to gender, ethnic group and school decile?" we used the parameter estimates from our main model (Table 4) to investigate whether the association of the gender, ethnic group and decile variables with PAT: Reading Comprehension in our model changed over time.

4. Findings

4.1. How does the variation in student reading achievement within schools compare with that between schools?

Our findings show that around 35 percent of the variance in PAT: Reading Comprehension scale score occurs at the school level, around 33 percent occurs at the student level and around 32 percent is residual (i.e. is not attributable to variation between schools or students). This is described in more detail in Section 9.1.

If we discount the residual variance—some of which will be associated with progress in PAT: Reading Comprehension scale score over time—then around 51 percent of the variance in PAT: Reading Comprehension scale score occurs at the school level (35 percent out of a total of 68 percent) and around 49 percent occurs at the student level (33 percent out of a total of 68 percent). This is a somewhat different picture to that described by the PISA study that has found that, for 15-year-olds, around a quarter of the variance in reading achievement occurs at the school level and around three-quarters of the variance occurs at the student level (OECD, 2010, Table II.5.1).

When trying to understand the difference in between-school and within-school variation between our findings and those of PISA, it is worth considering the following points:

1. The PISA data and the PAT: Reading Comprehension data are focussed on students of different age groups. While the PAT: Reading Comprehension data does contain some assessment records for 15-year-old students, these are only a small proportion of the total (assessment records for students in Year 10, some of whom will be 15 years old, make up around 5 percent of all assessment records). The majority are primary school students.
2. Because of the way that students from several contributing schools will go to the same intermediate school and students from several primary and intermediate schools will go to the same high school, school decile (which likely explains some of the variation in reading comprehension) is more variable for students in Years 4, 5 and 6 than it is for students in Years 7 and 8 where it is in turn more variable than it is for students in Years 9 and 10.
3. The statistical model that the PISA study uses to estimate variance at the school and student levels is different from the model we have used (and, in particular, does not incorporate multiple assessments of the same student).

4.2. Does student progress in reading occur at substantially different rates according to gender, ethnic group and school decile?

Before answering this question, it is worth noting that if our model was used to estimate the reading achievement of all New Zealand students in Year 4, Year 5 and Year 8, then the results of the reading achievement studies PIRLS and NMSSA (Chamberlain, 2014; Educational Assessment Research Unit & New Zealand Council for Educational Research, 2016) would be broadly replicated. In particular, for students in Year 4, our main model indicates that in reading comprehension:

- girls tend to have higher achievement than boys
- students at high decile schools tend to have higher achievement than students at low decile schools
- students who identify as New Zealand European tend to have the highest level of achievement in reading comprehension, followed in descending order of average achievement by students who identify as Asian, Other, Māori and Pasifika.

To now answer this section's question about student progress, our main model indicates that some groups of students make faster progress (boys, students at decile 1–2 schools, and students who identify with the Other ethnic group) or slower progress (students in decile 7–10 schools) than other students. This manifests in our model as statistically significant interactions between student and school characteristics and time (see Table 4 in Section 9).

However, the estimates in our model associated with these interactions are all quite small (Table 4). For example, the interaction of time with gender (which Table 4 shows having an estimated effect of 0.3149) means that our model indicates that boys make an average annual progress of 0.3149 *patc* points more than girls. This means that over 7 years (the period of time covered by the PAT: Reading Comprehension tests), the model indicates that boys make an average total progress of 2.2043 *patc* points more than girls. To contextualise this number, our data shows students making average annual progress of between 7 and 10 scale score points (see Table 3 in Section 9) and Darr et al. show that 2.2043 is well within the error of measurement of the PAT: Reading Comprehension tests (Darr et al., 2008). This last fact means the PAT: Reading Comprehension assessment can't reliably distinguish between the reading comprehension of any two students separated by 2.2043 *patc* points.

5. Discussion

5.1. Limitations

Any interpretation of our findings should recognise the limitations of our data and methods.

5.1.1. Data

The PAT: Reading Comprehension data is an administrative dataset that exists independent of our research. It is large, longitudinal and incorporates a robust measure of reading achievement. As such, it has provided us with a unique opportunity to address our research questions. However, it also limits the applicability of our findings.

Schools choose to use the PAT: Reading Comprehension tests and choose how to use them (including whether or not to administer them online or mark them electronically). These choices are likely to introduce biases into our data, and therefore limit our ability to generalise our findings. Because we are interested in progress, we only included students with four or more assessment records in our analyses—another choice that likely introduced bias.

Another consequence of the use of an administrative dataset to address our research questions is that we are only able to access limited information about students and schools (such as student demographic information and school decile). This has perhaps limited our ability to explain much of the variance in PAT: Reading Comprehension scale score at both the student and school levels.

5.1.2. Modelling

Any statistical model represents a compromise between data and narrative. Three modelling choices that we made epitomising this compromise are our decision not to use a Growth Mixture Model (e.g. Muthén, 2004), our decision not to model the 'summer reading slump' (Lai et al., 2009), and our decision not to model cohort effects (which, had we modelled them, would have allowed us to describe any differences in the patterns of achievement and progress between cohorts).

We did not use a Growth Mixture Model because we would have had to reframe our research questions to suit the modelling method, comparing instead the characteristics of groups of students classified only by the shape of their progress trajectories. We felt that this represented an unnecessarily complicated approach to answering our research questions.

We did not model the effect of the summer reading slump because there was limited evidence for this phenomenon in the distributions of reading achievement *across our entire dataset* (as evidenced by Table 3 in Section 9). This is not to deny the existence of the phenomenon, but rather to note that its effect is likely to be more evident in population subgroups than in the whole population. Consequently, modelling the effect of the summer reading slump across the whole population would require more complex interaction terms in our models. The way the summer reading slump will manifest in our main model as it stands is in the interactions between decile or ethnic group and rate of progress. However, it would be worth investigating the summer reading slump in the PAT: Reading Comprehension data more fully.

As for cohort effects, we did in fact include these in preliminary statistical models where they were statistically significant, but had very small parameter estimates. Given the magnitude of these cohort effects and the muddying impact they would have introduced to our answers to our research questions, we decided to omit them from our main model.

5.1.3. Results

When interpreting models such as the ones used in this research, it is important to remember that the models are detecting patterns of behaviour that are only evident when the data associated with *many people* is taken into consideration. As such, the models are unsuitable for predicting outcomes for any particular individual.

5.2. Implications

Our findings imply that such differences by gender, ethnic group or school decile in average PAT: Reading Comprehension score that already exist at Year 4 are largely unchanged by Year 10. As with studies such as PIRLS and NMSSA, this asks us what has happened in homes and schools prior to Year 4 that has led to such a pattern of differences. It also asks us what is happening in homes and schools between Year 4 and Year 10 that results in essentially parallel (average) progress for different population subgroups. Of course, this is not all bad—progress is occurring for all population subgroups, and there are no strong Matthew effects. However, it is not the equitable outcome to which many educational agencies are committed (e.g. Ministry of Education, 2015a, p. 17). Ultimately, these concerns highlight for us the importance of interventions designed to accelerate the reading progress of lower achieving population subgroups. One such intervention is described by Lai et al. (2009), and other interventions are summarised in the discussion section of the same paper.

Our model also has implications for the debate around progress through the National Standards in Reading. There is currently no national, longitudinal data recording student achievement against the National Standards. In the absence of such data, and together with thresholds on the PAT: Reading Comprehension scale associated with moving from one reporting band of the National Standards in Reading to another (e.g. Lawes, 2016, p. 13), the parameter estimates from our model (see Table 4) could be used to robustly simulate the percentages of students receiving various sequences of judgements against the National Standards. This simulation could then be used to evaluate the extent to which the National Standards in Reading are able to meaningfully describe progress (comparable to Lawes, 2016).

Part B: Technical detail

This part of the report provides the technical and methodological detail underpinning our research.

6. A measure of reading comprehension

The PAT: Reading Comprehension assessment is a sequence of standardised tests developed by the New Zealand Council for Educational Research (NZCER) specifically for use in New Zealand schools (Darr et al., 2008). The tests are designed to help classroom teachers to understand the achievement and progress of their students in reading comprehension as it is described in *The New Zealand Curriculum* (Ministry of Education, 2007) and, as such, are low-stakes.

PAT: Reading Comprehension tests cover reading comprehension achievement typically shown by students in Years 4 to 10. The tests are all linked to an equal interval scale known as the PAT: Reading Comprehension scale with units known as *patc* units.

Schools choose whether to purchase and use the PAT: Reading Comprehension assessment. They are supported by NZCER in their use of the PAT: Reading Comprehension assessment. In particular, NZCER provides advice about when, how and how often to administer the assessment, how it will be marked, and how to interpret and make use of the results. Tests can be adminstered on paper or online. Paper tests can be manually marked by the school, or scanned and electronically marked by NZCER. Online tests are marked electronically. It is the records of electronically marked assessments that make up the data described in this paper.

7. Data

7.1. Cleaning

The first stage of data cleaning involved fuzzy matching to link student records longitudinally. Much of the data was already longitudinally linked, but the linking variable (National Student Number—see Ministry of Education, 2015b) was used less extensively in the earlier years of data collection. Where National Student Number was missing, we uniquely identified each assessment record. Fuzzy matching then reduced the number of unique identifiers by around 7 percent.

The second phase of data cleaning involved ensuring that, for each student, their gender and ethnic group were recorded consistently over multiple assessments. Because of the size of the data, this could not be done manually. Students who had inconsistent gender records were assigned a final gender corresponding to the gender that was most commonly recorded (around 1.6 percent of all students at this stage of cleaning). Students who had an equal number of assessments with gender recorded as female and gender recorded as male were randomly assigned a final gender (around 0.3 percent of all students at this stage of cleaning). Students' final ethnic group identifications consisted of every ethnic group that they had ever identified with.

The third phase of data cleaning involved cleaning assessment records. Because the data collection period was long (7 years), we grouped assessments by half-year so as not to over-burden our analysis with too many time values. Where a student had been assessed more than once in a half-year, we took the average of the assessment results. Where a student had been assessed at more than one school in a half-year, we randomly selected the assessment to include in our data (around 0.2 percent of all students at this stage of cleaning).

The fourth phase involved removing assessment records for students who had been assessed fewer than four times. This reduced the size of our data substantially (from 864,632 to 352,473 assessment records), but was necessary to be able to statistically model reading progress in a valid way.

The resulting data consisted of 352,473 assessment records of 70,505 students at 716 schools, where each student had been assessed four or more times between 2008 and 2015 when they were in Years 4 through 10. Table 1 summarises the data.

Table 1 **Numbers of assessments, students and schools in the PAT: Reading Comprehension data**

Data feature	Number
Assessments	
2008	7,301
2009	18,790
2010	46,860
2011	66,682
2012	77,243
2013	73,836
2014	56,583
2015	5,178
Students	
Girls	35,836
Boys	34,669
New Zealand European	50,972
Māori	13,767
Pasifika	6,571
Asian	7,087
Other	6,603
Schools	
Deciles 1–2	100
Deciles 3–4	109
Deciles 5–6	128
Deciles 7–8	153
Deciles 9–10	226

Note that the assessment data from 2015 represents assessments sat at the very beginning of the school year.

We used the software environment R for all of our data cleaning and processing (R Core Team, 2015).

8. Analysis

Following an intial exploration of the data, we used multilevel linear models to summarise the features of the PAT: Reading Comprehension data and address our research questions. Multilevel linear models are applicable to data in which one unit of analysis is grouped within another—for example, when student data is grouped within school data. Multilevel models seek to specify the value of a dependent variable (such as reading achievement), based on the values of the independent variables (such as school characteristics and student characteristics) where some of the variables vary according to one unit of analysis (such as school characteristics) and other variables vary according to another unit of analysis (such as student characteristics). When making inferences from a statistical model of multilevel data, there is less chance of making a type I error (finding a relationship when one doesn't exist) than when using a single-level model. This is because single-level models tend to underestimate the variance that occurs at higher data levels—variance that multilevel models explicitly incorporate (e.g. Raudenbush & Bryk, 2002). Because our data is large and there is little risk of undercoverage, we did not use weighting or resampling methods.

Our research questions—particularly the first research question—required that our multilevel models would at least have 'random intercepts'. So-called random-intercept models account for variation of the base estimate of the dependent variable at one data level (the intercept) by the groupings at higher data levels. In our context this means that the models are required to account for the way that PAT: Reading Comprehension scale score (an assessment-level variable) varies by student and school. We could also have allowed our models to have 'random slopes'. So-called random-slope models account for the way the impact of an independent variable at one data level (a slope) varies across groups at higher data levels. We interpreted this to mean that a model would account for the way that the impact of time (an assessment level variable) on PAT: Reading Comprehension scale score varied by student and school. We explored the possibility of fitting random-slope models. However, we found that the variance of the slopes in each of the models we explored was small—in fact, smaller than the residual variance of the model. We therefore concluded that the effect of explanatory variables did not vary sufficiently by student or school to warrant the inclusion of a random slope in the model.

In addition to ruling out the inclusion of random slopes during model development, we also ruled out a number of variables at the school level that were not significantly associated with reading progress in the presence of student demographic factors and school decile. These were school roll, the percentages of students at a school who identifed with various ethnic groups and the rural or urban nature of a school. None of these variables were included in our final main model.

It is further worth noting that school decile is an ordered categorical variable with 10 categories. We grouped these into pairs (deciles 1 and 2 are grouped, etc.) to reduce the number of comparisons being made. We also fitted interactions with time which allowed us to explore how different student- and school-level factors are related to student progress.

In summary, we fitted two 3-level random intercept models to the PAT: Reading Comprehension data (see, for example, Finch et al., 2014). We refer to these as the null model and the main model. As multilevel models allow, in our main model, the intercept and slope parameters at the assessment level are considered variables at the student level (with the assessment-level intercept varying randomly), and the two intercepts

at the student level are considered as variables at the school level (with the one associated with the assessment-level intercept varying randomly).

For both of our models:

Y represents PAT: Reading Comprehension scale score.

X represents assessment time measured in year levels since the first half of Year 4.

Z_1 represents student identification as a boy.

Z_2 through Z_5 represent student identification with the Māori, Pasifika, Asian and Other ethnic groups respectively.

W_1 through W_4 represent school membership of decile 3 or 4, decile 5 or 6, decile 7 or 8, or decile 9 or 10 respectively.

e represents random error in the modelling of reading achievement at the assessment level.

u_0 represents random error in the modelling of reading achievement at the student level.

v_{00} represents random error in the modelling of reading achievement at the school level.

For clarity, we suppress the traditional representation of: assessment-level variables with an additional subscript of ijk (they vary by assessment, student and school); assessment-level parameters and student-level variables with an additional subscript of jk (they vary by student and school); and student-level parameters and school-level variables with an additional subscript of k (they vary by school).

We used the software environment R for all of our statistical analysis and, in particular, for multilevel modelling we used the R package 'lme4' developed by Bates et al. (2015) and described in Finch et al. (2014).

8.1. Analysis to compare the variation in student reading achievement within schools with that between schools

We used a '3-level null model' to answer our first research question—that is, a 3-level model with no independent variables. At the assessment level it has equation:

$$Y = \beta_0 + e$$

At the student level our null model has equation:

$$\beta_0 = \gamma_{00} + u_0$$

At the school level our null model has equation:

$$\gamma_{00} = \delta_{000} + v_{00}$$

Here:

β_0 represents the average PAT: Reading Comprehension scale score for a student.

γ_{00} represents the average PAT: Reading Comprehension scale score for a school.

δ_{000} represents the average PAT: Reading Comprehension scale score for all students in all schools.

Of paricular interest in this model are the variances of u_0 and v_{00}. Their relative magnitudes will provide us with evidence about how the variation in student reading achievement over time compares with the variation within a school and the variation between schools. Our modelling process will provide us with estimates of these parameters.

8.2. Analysis to determine the extent to which student progress in reading differs according to gender, ethnic group and school decile

Exploratory analyses suggested that progress as shown in the PAT: Reading Comprehension data was linear. This manifests in Table 3 and Figure 3. Therefore, prior to fitting to the data, at the assessment level the model we used to address our second research question had equation:

$$Y = \beta_0 + \beta_1 X + e$$

At the student level our model had equations:

$$\beta_0 = \gamma_{00} + \gamma_{01}Z_1 + \gamma_{02}Z_2 + \gamma_{03}Z_3 + \gamma_{04}Z_4 + \gamma_{05}Z_5 + \gamma_{06}Z_1Z_2 + \gamma_{07}Z_1Z_3 + \gamma_{08}Z_1Z_4 + \gamma_{09}Z_1Z_4 + u_0$$

$$\beta_1 = \gamma_{10} + \gamma_{11k}Z_1 + \gamma_{12}Z_2 + \gamma_{13}Z_3 + \gamma_{14}Z_4 + \gamma_{15}Z_5$$

At the school level our model had equations:

$$\gamma_{00} = \delta_{000} + \delta_{001}W_1 + \delta_{002}W_2 + \delta_{003}W_3 + \delta_{004}W_4 + v_{00}$$

$$\gamma_{10} = \delta_{100} + \delta_{101}W_1 + \delta_{102}W_2 + \delta_{103}W_3 + \delta_{104}W_4$$

Here:

β_0 represents the average PAT: Reading Comprehension scale score of a student in the first half of Year 4.

β_1 represents the average rate of progress of a student up the PAT: Reading Comprehension scale.

γ_{00} represents the average PAT: Reading Comprehension scale score of a school at the beginning of Year 4.

$\gamma_{01}, \gamma_{02}, \gamma_{03}, \gamma_{04}, \gamma_{05}$ represent the average differences in PAT: Reading Comprehension scale score associated with student identification as a boy, Māori, Pasfika, Asian or Other ethnic group respectively.

$\gamma_{06}, \gamma_{07}, \gamma_{08}, \gamma_{09}$ represent the average differences in PAT: Reading Comprehension scale score associated with student identification as a boy who is Māori, Pasfika, Asian or from the Other ethnic group category respectively. That is, these parameters represent interactions between

student identification as a boy and student identification as Māori, Pasfika, Asian or Other ethnic group.

γ_{10} represents the average rate of progress of a school up the PAT: Reading Comprehension scale.

$\gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{14}, \gamma_{15}$ represent the average differences in the rate of progress up the PAT: Reading Comprehension scale associated with student identification as a boy, Māori, Pasfika, Asian or Other ethnic group respectively.

δ_{000} represents the average PAT: Reading Comprehension scale score of all students in all schools at the beginning of Year 4.

$\delta_{001}, \delta_{002}, \delta_{003}, \delta_{004}$ represent the average differences in PAT: Reading Comprehension scale score associated with school identification as being in decile 3 or 4, decile 5 or 6, decile 7 or 8, and decile 9 or 10 respectively.

δ_{100} represents the average rate of progress of all students in all schools along the PAT: Reading Comprehension scale.

$\delta_{101}, \delta_{102}, \delta_{103}, \delta_{104}$ represent the average differences in the rate of progress of students along the PAT: Reading Comprehension scale associated with school identification as being in decile 3 or 4, decile 5 or 6, decile 7 or 8, and decile 9 or 10 respectively.

Of particular interest in this model are the student-level equation for β_1 and the school-level equation for γ_{10}. The strength and magnitude of association of the variables in these equations will provide us with evidence regarding the extent to which student progress in reading differs according to gender, ethnic group and school decile.

9. Results

9.1. Comparing the variation in student reading achievement within schools with that between schools

Parameter estimates for the null model described above are shown in Table 2.

Table 2 **Parameter estimates for the null model**

Effect	Estimate	SE	
Fixed effects			
Intercept	51.2047	0.4229	***
Random effects			
School-level variance (i.e. the variance of v_{00})	117.0		
Student-level variance (i.e. the variance of u_0)	111.4		
Residual variance (i.e. the variance of e)	108.7		

Table 2 tells us that the overall average score for all of the assessments in the PAT: Reading Comprehension data (notated δ_{000}) was around 51 *patc*. Table 2 also tells us that the total variance for all of the assessments in the PAT: Reading Comprehension data was the sum of: 117 *patc* (around 35 percent of the total variance); 111.4 *patc* (around 33 percent of the total variance); and 108.7 *patc* (around 32 percent of the total variance). The $-2\log\text{Likelihood}$ of the null model was 2781882 with 4 degrees of freedom. This essentially means that the null model describes the variation in the data well.

9.2. The extent to which student progress in reading differs according to gender, ethnic group and school decile

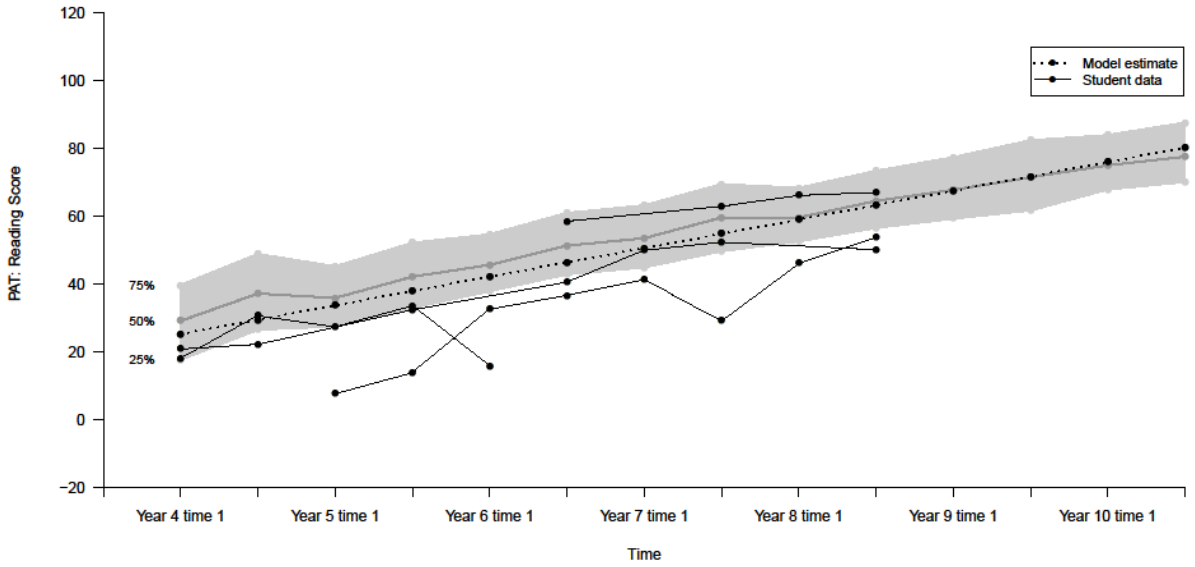
Exploratory analyses suggested that progress as shown in the PAT: Reading Comprehension data was linear. This is indicated in Table 3 (showing distributions of reading achievement by time).

Table 3 **The distribution of PAT: Reading Comprehension scale scores by time**

Year level	Time	Number of assessments	Mean assessment scale score	SD assessment scale score
Year 4	1	28,735	32.33	15.54
	2	13,423	38.73	15.85
Year 5	1	37,581	39.33	12.92
	2	19,653	43.97	13.87
Year 6	1	39,635	48.99	12.4
	2	21,183	53.38	13.53
Year 7	1	49,534	57.19	12.9
	2	29,304	61.88	14.05
Year 8	1	42,056	63.24	11.41
	2	28,706	67.55	12.47
Year 9	1	17,959	73.38	12.71
	2	7,426	77.2	13.81
Year 10	1	12,426	80.52	11.07
	2	4,852	82	11.88

Figure 3 shows the progress of some randomly selected students displayed against a backgroud showing estimates of national achievement norms recovered from the PAT: Reading Comprehension data.

Figure 3 **The distribution of PAT: Reading Comprehension scale scores by time**



PAT: Reading Comprehension scale scores (in grey) are shown together with the progress trajectories of four randomly selected male, New Zealand European students from decile 1 or 2 schools (in black), and the model estimated progress trajectory (dotted line). The distributions of PAT: Reading Comprehension scale scores were estimated using resampling methods to more accurately represent national norms. This is in contrast with Table 3 where the characteristics of the data are of interest, and no resampling is used.

Figure 3 demonstrates that, while it is reasonable to model reading progress linearly across a population, not all individual students' progress trajectories will be linear. This deviation of individual progress from the modelled linear progress contributes to the residual variance reported in Table 2 and Table 4. Table 4 displays the parameter estimates of our main model. It is worth noting that, for the most part, we only included effects in the model that were statistically significant. The exception to this, for ease of interpretation, is the inclusion of all interactions associated with decile. The $-2\log\text{Likelihood}$ of the main model was 2496941 with 30 degrees of freedom. This essentially means that the main model describes our data well, and certainly better than the null model.

Table 4 **Parameter estimates for the main model**

Effect	Estimate	SE	
Fixed effects			
Intercept	29.2	0.358	***
Assessment level			
Time	8.458	0.04143	***
Student level			
Gender (boys)	-4.221	0.1008	***
Māori	-3.878	0.2576	***
Pasifika	-4.568	0.3145	***
Asian	-0.318	0.1403	*
Other	-0.935	0.167	***
Gender*Pasifika	0.9365	0.2758	***
School level			
Deciles 3–4	3.27	0.4468	***
Deciles 5–6	5.114	0.4306	***
Deciles 7–8	7.118	0.4115	***
Deciles 9–10	8.744	0.3917	***
Cross-level interactions			
Time*Gender	0.3149	0.01951	***
Time*Pasifika	-0.0881	0.03696	*
Time*Other	0.1338	0.03085	***
Māori*Deciles 3–4	-0.8298	0.2922	**
Māori*Deciles 5–6	-0.7099	0.2984	*
Māori*Deciles 7–8	-0.6657	0.2906	*
Māori*Deciles 9–10	-0.2096	0.3018	
Pasifika*Deciles 3–4	-0.5103	0.334	
Pasifika*Deciles 5–6	-1.884	0.3834	***
Pasifika*Deciles 7–8	-1.856	0.3476	***
Pasifika*Deciles 9–10	-2.197	0.3616	***
Time*Deciles 3–4	-0.3264	0.05454	***
Time*Deciles 5–6	-0.5336	0.05098	***
Time*Deciles 7–8	-0.6533	0.04643	***
Time*Deciles 9–10	-0.6864	0.04368	***
Random effects			
School-level variance	4.39		
Student-level variance	103.44		
Residual variance	41.59		

*p < 0.05. **p < 0.01. ***p < 0.001.

As an example of how to interpret the estimates in Table 4, we use the model to determine the average PAT: Reading Comprehension scale score for a Pasifika boy in a decile 3 school at the beginning of Year 4. Table 5 shows the part of Table 4 that we use for this purpose. Essentially we have dropped from Table 4 anything related to time (in this example we are interested in achievement at the beginning of Year 4, not progress), student descriptors unrelated to being a Pasifika boy and school descriptors unrelated to being in a decile 3–4 school.

Table 5 **Modelling the average PAT: Reading Comprehension scale score for a Pasifika boy in a decile 3 school at the beginning of Year 4**

Effect	Estimate	SE	
Fixed effects			
Intercept	29.2	0.358	***
Student level			
Gender (boys)	-4.221	0.1008	***
Pasifika	-4.568	0.3145	***
Gender*Pasifika	0.9365	0.2758	***
School level			
Deciles 3–4	3.27	0.4468	***
Cross-level interactions			
Pasifika*Deciles 3–4	-0.5103	0.334	

*p < 0.05. **p < 0.01. ***p < 0.001.

To obtain the average PAT: Reading Comprehension scale score for a Pasifika boy in a decile 3 school at the beginning of Year 4, we simply add the column of estimates in Table 5 to obtain 24.1072 *patc* points.

To determine the average annual progress of the same student, we would use the part of Table 4 that is shown as Table 6. Essentially we have dropped from Table 4 anything *not* related to time (as we are now interested in progress), and how it interacts with the student descriptors for being a Pasifika boy, and the school descriptors for being in a decile 3–4 school.

Table 6 **Modelling the average annual progress for a Pasifika boy in a decile 3 school**

Effect	Estimate	SE	
Fixed effects			
Assessment level			
Time	8.458	0.04143	***
Cross-level interactions			
Time*Gender	0.3149	0.01951	***
Time*Pasifika	-0.0881	0.03696	*
Time*Deciles 3–4	-0.3264	0.05454	***

*p < 0.05. **p < 0.01. ***p < 0.001.

To obtain the average annual progress of a Pasifika boy in a decile 3 school, we simply add the column of estimates in Table 6 to obtain 8.3584 *patc* points per year.

Table 4 also tells us that, after accounting for all of the variables in the model, the total variance for all of the assessments in the PAT: Reading Comprehension data was the sum of: 4.39 (the variance at the school level—that is, the variance of v_{00}); 103.44 (the variance at the student level—that is, the variance of u_0); and 41.59 (the residual variance—that is, the variance of e).

References

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

Chamberlain, M. (2014). *PIRLS 2010/11 in New Zealand: An overview of findings from the third cycle of the Progress in International Reading Literacy Study (PIRLS)* (revised ed.). Wellington: Ministry of Education. Retrieved from https://www.educationcounts.govt.nz

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

Educational Assessment Research Unit & New Zealand Council for Educational Research. (2016). *English: Reading 2014—overview*. Wellington: Ministry of Education.

Finch, W. H., Bolin, J. E., & Kelley, K. (2014). *Multilevel modelling using R*. Boca Raton, FL: Chapman & Hall/CRC.

Gilmore, A., & Smith, J. (2011). *NEMP report 53: Writing, reading and mathematics report 2010*. Dunedin: University of Otago, Educational Assessment Research Unit.

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.

Lawes, E. (2016). *Using PAT: Mathematics to simulate student progress through the National Standards.* Wellington: New Zealand Council for Educational Research.

Ministry of Education. (2007). *The New Zealand curriculum: For English-medium teaching and learning in years 1–13*. Wellington: Learning Media.

Ministry of Education. (2010a). *Designing the literacy learning progressions*. Wellington: Learning Media. Retrieved from http://www.literacyprogressions.tki.org.nz/Background

Ministry of Education. (2010b). *Designing the reading and writing standards for years 1–8*. Wellington: Learning Media. Retrieved from http://nzcurriculum.tki.org.nz/National-Standards/Key-information/Fact-sheets/Background-papers

Ministry of Education. (2010c). *The literacy learning progressions*. Wellington: Learning Media. Retrieved from http://www.literacyprogressions.tki.org.nz/

Ministry of Education. (2013a). *The national administration guidelines (NAGs): National administration guideline 2A.* Wellington: Author. Retrieved from http://www.education.govt.nz/ministry-ofeducation/legislation/nags/#NAG2A

Ministry of Education. (2013b). *Reading and writing standards*. Wellington: Author. Retrieved from http://nzcurriculum.tki.org.nz/National-Standards/Reading-and-writing-standards

Ministry of Education. (2015a). *Four year plan 2015–2019*. Wellington: Author. Retrieved from http://www.education.govt.nz/ministry-of-education/publications/four-year-plan-and-statements-of-intent/four-year-plan-2015-2019/

Ministry of Education. (2015b). *National student index (NSI)*. Wellington: Author. Retrieved from https://nsi.education.govt.nz/

Ministry of Education. (2016). *Programme for the International Assessment of Adult Competencies (PIAAC). International Survey of Adult Skills (ISAS)*. Wellington: Author Retrieved from http://www.educationcounts.govt.nz/data-services/data-collections/international/piaac

Muthén, B. (2004). Growth mixture modelling. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 345–368). Thousand Oaks, CA: Sage.

Nusche, D., Laveault, D., MacBeath, J., & Santiago, P. (2012). *OECD reviews of evaluation and assessment in education: New Zealand 2011*. Paris: OECD. Retrieved from http://dx.doi.org/10.1787/9789264116917-en

OECD. (2010). *PISA 2009 results: Overcoming social background—equity in learning opportunities and outcomes (volume II)*. Paris: OECD. Retrieved from http://dx.doi.org/10.1787/9789264091504-en

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Research Division, Ministry of Education. (2014). *Annual monitoring of reading recovery*. Wellington: Ministry of Education. Retrieved from https://www.educationcounts.govt.nz/publications/series/1547

Robinson, V., Hohepa, M., & Lloyd, C. (2015). *School leadership and student outcomes: Identifying what works and why* (redacted version). Wellington: Ministry of Education.

Soler, J., & Openshaw, R. (2006). *Literacy crises and reading policies: Children still can't read!* New York: Routledge.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360–407. doi:10.1598/RRQ.21.4.1

Satherley, P., & Lawes, E. (2007). *The Adult Literacy and Life Skills (ALL) survey: An introduction*. Wellington: Ministry of Education. Retrieved from http://www.educationcounts.govt.nz/publications/series/ALL

Telford, M., & May, S. (2010). *PISA 2009: Our 21st century learners at age 15*. Wellington: Ministry of Education. Retrieved from https://www.educationcounts.govt.nz/topics/research/pisa

Tunmer, W. E., Chapman, J. W., Greaney, K. T., Prochnow, J. E., & Arrow, A. W. (2013). Why the New Zealand national literacy strategy has failed and what can be done about it: Evidence from the Progress in International Reading Literacy Study (PIRLS) 2011 and reading recovery monitoring reports. *Australian Journal of Learning Difficulties*, *18*(2), 139–180.

Walker, M., Udy, K., & Pole, N. (1996). Adult literacy in New Zealand: Results from the International Adult Literacy Survey. Wellington: Ministry of Education. Retrieved from https://www.educationcounts.govt.nz/publications/literacy/5731