Literacy Professional Development Project

Contents

Executive summary, including the key findings

Outline of this report

The Literacy Professional Development Project began working with its first cohort of schools in February 2004. This report is based on the national data collected for those schools from then until they completed their involvement nearly two years later in November 2005.

The project has four outcomes:

- Evidence of improved student achievement
- Evidence of improved teacher content knowledge
- Evidence of improved transfer of understanding of literacy pedagogy to practice
- Evidence of professional learning communities.

The project uses particular needs analysis tools based on these outcomes to engage schools in an inquiry into the effectiveness of school-wide and classroom-based literacy practices. This report focuses on the first outcome, "evidence of improved student achievement". Further reports about the shifts in the other three outcomes will follow in May 2006.

The report begins by describing the project and the theory of improvement on which it is based. The next section describes the ninety-one schools that participated. Of these, forty-nine schools (593.6 teachers) had a writing focus and forty-two schools had a reading focus (589.5 teachers). The third section describes the tools, their limitations, and the limitations in the way they were used. This section includes a description of the types of analysis that has occurred with the data so far. The fourth section describes who is represented in each data set. These first four sections set the scene for the actual analysis and interpretation of the data in section five. The project-wide shifts in student achievement are explored through a range of different lenses. This is summarised below, with the improvements identified first, followed by the aspects of the initial findings that require further inquiry, and finally, the project's response to these aspects as we move into working with a new cohort of schools in 2006.

Key findings

Reading achievement data (40 schools/3787 students)

Improvements

The project chose *Supplementary Tests of Achievement in Reading* (STAR) as the tool it would use to monitor the shifts in reading comprehension. STAR is a standardised test that uses "the scores of a large representative sample of students to establish stanine norms for each class level" (NZCER, 2001, page 18)[1]. With stanines, it is assumed that if a student were to maintain their current rate of progress (with general teaching and

[1] NZCER (2001). *STAR Supplementary tests of achievement in reading years 4-9*. Wellington: NZCER.

student development) there would be no shift in a group of students' individual stanine scores over time and, therefore, no improvement in their mean stanine scores. Any shift seen in a year group's mean stanines show an altered trajectory and, in this case, the change is hypothesised to be attributed to increased teaching effectiveness.

The project has a focus on the "tail", as identified in national and international testing.

Whole cohort shifts

- The effect size of the shift in achievement for the whole cohort was 0.87, and the effect size of the shift in achievement for the lowest 23% of students at Time 1 was 1.97.
- Across all stanines and all year groups, there was a mean stanine shift of 0.56. The students in stanines 1–3 (the lowest 23% of students) at Time 1 had an average mean shift of 1.11.

Sub-group shifts

- A number of subgroups had a larger mean shift than this:
  - The students in years 4 and 7 had a stanine shift of 0.7.
  - The mean shift for Māori students was 0.59 (but on average they are 1 stanine lower than NZ European students).
  - The mean shift for Asian students was 0.68 (but on average they are 0.85 of a stanine lower than NZ European students).
  - Boys had a mean shift of 0.6 (with a mean stanine at Time 3 equivalent to the mean stanine for girls at Time 1).
- Māori students had a greater difference in mean raw scores for all year groups when compared to the whole year cohort except year 6.
- A number of subgroups had proportionally fewer students in the stanine 1–3 band than the STAR national picture by Time 3 (STAR norms = 23%):
  - The percentage of Māori students in the stanine 1–3 band was 30% at Time 1 and 18.8% at Time 3
  - The percentage of Asian students in the stanine 1–3 band was 32% at Time 1 and 18.3% at Time 3
  - The percentage of boys in the stanine 1–3 band was 28% at Time 1 and 17.4% at Time 3.
- The percentage of Pasifika students in the stanine 1–3 band was 40.5% at Time 1 and 26% at Time 3.
- Forty percent of stanine 1 students made a significant shift (2+ stanine shift), and 34% of stanine 2 students made a significant shift.

Aspects of the analysis of this data that require further inquiry

- The Pasifika students' mean stanine shift is less than the cohort as a whole (0.53 compared with 0.56) and both their Time 1 and Time 3 mean stanine score was the lowest for all sub-groups.
- A large proportion of students in stanine 1 at Time 1 are still within the stanine 1–3 band at Time 3.
  - Within the whole cohort, 84% of the students in stanine 1 at Time 1 are still within this band at Time 3.
  - In two subgroups, the percentage of students still within the stanine 1–3 band at Time 3 exceeds 84%. They are:
    - Boys: 88% of the students in stanine 1 at Time 1 are still within this band
    - Pasifika students: 95% of the students in stanine 1 at Time 1 are still within this band.
- It appears that there is a ceiling effect with STAR that impacts on the project's ability to describe the improvements made over the two years by students in stanines 7–9 and / or year 6.
- Year 6 data seems to often be different from the other year groups in all sorts of ways, for example:
  - The group has the lowest overall mean stanine shift, which may be due to a ceiling effect, but the largest mean stanine shift for the students in the lowest 3 stanines.
  - The subtest raw score difference for the Māori student cohort is less than that for the cohort as a whole, and yet is greater for every other year group.

Project response

- The project has decided to use asTTle: Reading for monitoring reading comprehension in 2006 for a number of reasons including that there does not appear to be a ceiling effect with this tool.
- Material will be sourced to inform the facilitation, leadership, and teaching and learning about Pasifika students and literacy teaching and learning.

Writing achievement data (45 schools/1063 students)

Improvements

The project chose asTTle: Writing as the tool that it would use to monitor the shifts in writing achievement. asTTle is a criterion-based test that is based an assumption that if a student were to maintain their current rate of progress (with general teaching and student development) there would be a shift of 1 curriculum sub-level over two years. The data presented is from a representative sample of approximately 10%. In each

school the facilitator moderated this random sample of students and used it to help the teachers to understand the asTTle marking schedules. The effect of using only the facilitator-moderated data when teacher content knowledge was unknown was to increase the likely accuracy and reliability of the data for the project at Time 1.

The project has a focus on the "tail", as identified in national and international testing.

Whole cohort shifts

- The effect size of the shift in achievement for the whole cohort was 1.30, and the effect size of the shift in achievement for the lowest 20% of students at Time 1 was 2.05.
- The whole cohort had a mean gain of 2.5 curriculum sub-levels over the two years (compared with the national mean gain of 1 curriculum sub-level over two years), and the lowest 20% of students at Time 1 gained 4 curriculum sub-levels.
- Years 4, 5, and 6 (at Time 3) were more than 1 curriculum sub-level above the national picture. Year 7 and 8 (at Time 3) was more than 2 curriculum sub-levels above the national picture.
- The lowest 20% of students in years 4–6 (at Time 3) are less than 1 curriculum sub-level below the mean curriculum sub-level for the national picture.

Sub-group shifts

- The lowest 20% of each year group of Māori students has a greater achievement shift than each year group as a whole.

Aspects of the analysis of this data that require further inquiry

- Years 6 and 7 mean scores for the students in the lowest 20% are not much higher than years 4 and 5 mean scores for the students in the lowest.
- The lowest 20% of students in years 7 and 8 (at Time 3) were more than 1 curriculum sub-level below the national picture.
- There are more boys in the lowest 20% of students and their pace of achievement is less than the girls in this group.
- The Pasifika cohort was not large enough to analyse shifts for each year group, so the project is unsure whether the shifts in writing are similar to the shifts (or lack of shift) in reading.

Project response

- asTTle: Writing scores will be collected from all students in the 2006 cohort. The facilitators will guide schools through the asTTle moderation process as detailed on TKI. If the process is followed, we should be able to trust the data and accept as reasonable a degree of variability in the marking consistent with that obtained in the large scale marking exercise associated with asTTle norming. With greater student numbers and with three data collection points, we will be able to make more sense of what various sub-groups of students are achieving.

Comparison between reading and writing

It appears that the effect sizes in writing for years 5 and 7 are greater than those in reading. The project will have a stronger reading-writing link for the February 2006 schools. The project will be able to monitor the difference between reading and writing within schools as well as between schools.

A description of the professional development project

The project is based on the premise that effective classroom teaching will lead to improved student achievement. The focus of the contract was to provide the schools with a professional development programme that brought about improvements in students' learning and achievement in literacy, teachers' content knowledge, and teachers' understanding of literacy pedagogy and teacher practice.

Another key premise is that professional learning is best done on site with other teachers. It was expected that teachers would learn within a professional learning community so that the learning and improvement culture would continue after each school had completed the two years on the project. It was thought that these communities would initially be developed by the facilitator, who would work as a "visiting leader", and that over time school leadership – principals, literacy leaders, syndicate leaders – would lead the professional learning. There was an expectation that these changes at both the classroom and the school-wide level would take time and that schools would go through the changes at different paces. The first cohort started in February 2004 and finished in November 2005.

Project outcomes

The project has four outcomes:

- Evidence of improved student achievement
- Evidence of improved teacher content knowledge
- Evidence of improved transfer of understanding of literacy pedagogy to practice
- Evidence of professional learning communities.

National needs analysis tools were developed to help the schools and facilitators inquire into the effectiveness of school practices and student achievement. The tools had three purposes:

- To engage the schools in a process of inquiry and evidence-based conversations
- To build a rich description of practice within a school to help prioritise the professional learning needs
- To identify shifts in the outcomes for individual schools and for the project as a whole.

The tools used to identify evidence of improved student achievement were:

- Writing: asTTle writing years 4–8
- Reading comprehension: STAR years 3–8

Schools were also supported to use other tools for particular monitoring and / or diagnostic tasks (for example, observation surveys and running records for the ongoing monitoring of effective classroom teaching for particular students).

Project details

The project:

- follows an evidence-based inquiry model with measurable improved student achievement as the primary goal;
- develops/enhances a strong professional learning community focused on quality teaching that is informed by achievement data, which leads to improved student achievement.

Schools chose either a reading comprehension or a writing focus while in the project.

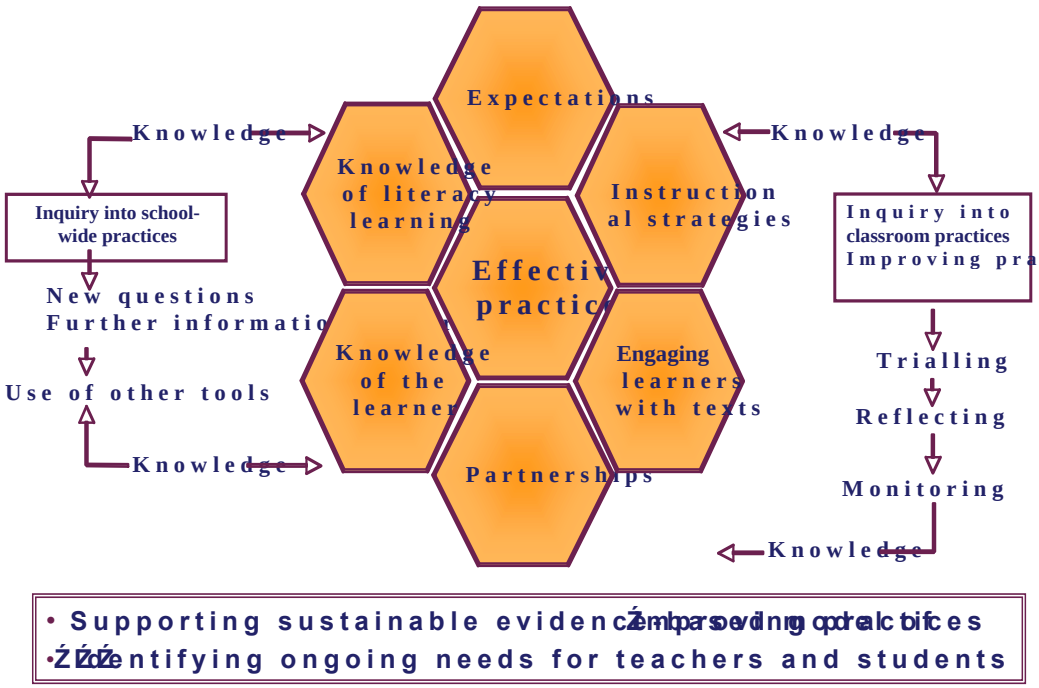
The project has three phases:

- **Phase 1**: An inquiry into learning – building an informed knowledge, evidence, and professional learning base
- **Phase 2**: Building knowledge and implementing change through active learning
- **Phase 3**: Evaluating and sustaining change.

An evaluation tool was used to monitor schools' progression through the phases.

The diagram below shows Phase 1 on the left-hand side, Phase 2 on the right-hand side, and Phase 3 as a continuous cycle of school-wide and classroom-based inquiry into the effectiveness of practice.

Figure 1: The three phases of the Literacy Professional Development Project



Description of school visits

The facilitators kept a record of schools visits over the two years the schools were in the project. This gave the project information about the purpose of the visits, who was visited, and what actually occurred. The facilitators' notes for each visit also identified the current issues and the next steps of the learning journey for each school.

The length of time for each visit ranged from half a day to two days, the visits had a range of purposes, and the facilitators worked with a variety of people at each visit. Table 1 shows that 40% of the visits included time with the principal, literacy leader/s, and senior management. The rest of the time was spent with teachers at an individual, group, or whole school level.

Table 1: Description of who was visited by facilitators

Description of who was visited	*% of time with each*
Principal and/or literacy leader/s	35%
Senior management	5%
Syndicate meetings	5%
Whole staff	22%
Individual teachers	20%
Other	13%[2]

The main purposes for the visits appear to be:

- Teacher literacy knowledge (18%)
- Assessment tool discussion (15%)
- Using a needs analysis tool (13%)
- Modelling the use of data to inform teaching (13%)
- Classroom observation (10%).

Changes to the project

The project length was initially to be one year. By July 2004, 77% of the first cohort of schools was underway or just beginning Phase 1. This evidence from facilitator evaluations, along with evidence from the research team about the progress in the research schools, was presented to the Minister of Education, and he agreed that all participating schools should be entitled to two years with the project. All schools were then invited to continue with the project in 2005. Thirty-three schools were unable to take advantage of the second year offered for a variety of reasons and left after the first year. These schools were described in the "Discontinuing Schools Report" in Milestone 1, 2005.

Facilitator practice

Facilitator practice is critical in enabling and supporting schools to examine the effectiveness of their classroom practice and improve student outcomes, teacher knowledge and practice, and leadership and professional learning communities. To support facilitator learning and inquiry into the effectiveness of facilitator practice, the project has a regional structure with experienced regional team leaders. Each leader works with their regional team at the four four-day national seminars held annually and the one-day regional meetings held three-weekly. The leaders meet six-weekly, as part of the Leadership and Effectiveness Team (LET), to discuss the ongoing

[2] Three facilitators appear to have an above average use of "other". It appears that this is because it is not de-selected, as it is the default choice.

monitoring of the effectiveness of the project and to make improvements to project and facilitator practice in response to this evidence.

A description of the schools involved

Ninety-one schools have completed two school years with the project. This particular cohort has been called "February 2004 cohort schools" in previous milestones. The following tables give some background information about these schools.

Table 2: Number and percentage of schools with a reading comprehension or writing focus and number and proportion of types of schools within the February 2004 cohort

	Number of schools *Total = 91*	*Percentage of total*
Reading focus	42	46
Writing focus	49	54
All	**91**	**100**
Full primary	40	45
Contributing	33	36
Intermediate	13	14
Years 7–15	2	2
Special	2	2
Restricted	1	1
All	**91**	**100**

Table 3: Decile ratings for the schools within the February 2004 cohort

School decile	*Number of schools* Total = 91	*Percentage of total*
1	4	4.4
2	13	14.3
3	7	7.7
4	14	15.4
5	5	5.5
6	12	13.2
7	10	11.0
8	11	12.1
9	10	11.0
10	5	5.5
All	**91**	**100**

Table 4: Size of the reading and writing schools within the February 2004 cohort

Number of reading teachers in each school	*Number of reading schools* Total = 42	*Number of writing schools* Total = 49
0 - 5	7	9
6 - 10	16	18
11 - 20	7	15
21 - 30	10	4
31+	2	3
All	**42**	**49**

The schools with 31+ teachers were all intermediates.

An explanation of the purposes for the data analysis from a project viewpoint

This project is a learning project. It is based on several assumptions.

One assumption is that learning to improve practice is about building knowledge and developing the abilities to know when one's practice is not effective and to do something about it. In other words, it is about developing an evidence-based inquiry habit of mind.

Another core assumption is that learning to improve practice involves other professionals who are genuinely motivated to understand their own and others' practice and take shared responsibility in improving practice. The role of leadership is to enable others to be more effective in their practice. That is, leaders are expected to provide the necessary structures, including those that motivate others to learn.

The project is also based on the assumption that what works to improve teachers' practice also works for leaders, facilitators, and the project as a whole. This means that people working at all levels of the project need to develop the habit of conducting an ongoing evidence-based inquiry into the effectiveness of their own practice, with the goal of enabling others to be more effective in their practice. Project leaders work to enable facilitators to be more effective in their practice, facilitators work to enable school leaders and teachers to be more effective in their practice, and in turn, school leaders and teachers work to enable students to be more effective in their practice.

In the earlier section, "A description of the professional development project", the three purposes for the needs analysis tools were described: to engage schools in an inquiry process, to build a rich picture that would inform the professional learning, and to identify the shifts in the outcomes over time. Given the set of assumptions described above, it would be expected that there would be three very similar purposes for the data analysis from a project level. An additional purpose supports the leadership role of the project developers.

The purposes of the needs analysis and interpretations at a project level are:

- to engage in a "sense-making" analysis in order to understand the efficacy and limitations of the needs analysis tools and the way they were used;
- to build a rich description of practices within the project to help prioritise the professional learning needs of facilitators and schools and to make informed changes to the design of the project;
- to engage others in this process of inquiry and evidence-based conversations;
- to identify shifts in the outcomes for the project as a whole.

Each of these purposes is discussed in more detail below.

Engaging in a "sense-making" analysis

The project needs to engage in a "sense-making" analysis in order to understand the efficacy and limitations of the tools and the way they were used.

Each of the tools used in the project has a range of limitations in the data that they provide. This set of limitations is inherent in each tool. Each tool has another set of limitations in the reliability and/validity in the data because of the way it was used. Some of these limitations in the tools and the way they were used were known before they were used. The Leadership and Effectiveness Team (LET) decided that the data provided would still be very useful in meeting the purposes of the needs analysis tools at both a school and project level. These limitations are described in the following section.

Other limitations – both within the tools and in the way they were used – have only become apparent during the analysis of the project data. Some are described in detail within the "Evidence-based analysis" section of this report. Others have been described in earlier milestones once the evidence of the limitations became apparent. This evidence has come from the case study evaluations, the discontinuing schools analysis, and from the embedded research reports. They include:

- the focus on learning intentions after completing the scenario;
- the lack of clarity around leadership goals, as leaders are not mentioned in the four project outcomes;
- the focus on data for the national picture rather than for building a school picture.

There are other limitations around the actual data analysis. The project has produced a lot of data or evidence. As discussed above, this varies in its level of trustworthiness. As Shulman[3] reminds us, "We often have lots of evidence to choose from; the problem is making sense of it and drawing the right lessons". We have had to make decisions about what evidence to use – what to notice and what to ignore – but have tried to make it obvious to the readers what we have looked for and why. The "why" has a lot to do with the second and third purposes for this analysis – the drive to model an inquiry into the effectiveness of project practices. We also have our own limitations around our understanding of data analysis and statistical processes and have sought help with this work.

In this report the evidence for the outcome, "evidence of improved student achievement", is analysed and interpreted through certain lenses. There is evidence of improvement but the analysis also raises some concerns, which generally require further inquiry. They also require some action or response by the project, which will be reviewed in the future. In this way, this report models the ongoing inquiry.

[3] Shulman L. (2005). *Seek Simplicity … and Distrust It*. Carnegie Foundation for the Advancement of Teaching. Downloaded 20 January 2005.

Building a rich description of practices

The project needs to build a rich description of practices within the project to help prioritise the professional learning needs of facilitators and schools and to make informed changes to the design of the project.

Through this inquiry, we hope to understand a lot more about the project than we did before undertaking this analysis. The four project outcomes are linked, as shown in the adapted diagram below (ERO, 2003)[4]. Figure 2. Effective facilitation enables school leaders and classroom teachers to be effective. There has been much thought and work put into finding out what makes for effective facilitation and how the project can support facilitator learning.

The concept of first and second order change described by Waters, Marzano, and McNulty (2003)[5] has been very useful in helping us to analyse the evidence about the shifts within the outcomes. The easier, or "first order", shifts seem to have been those associated with improvements of practice. The harder, or "second order", shifts appear to be those that require an examination of personal beliefs and a new way of working. Waters et al. have identified that communication needs to be different for a second order change than for a first order change. This suggests that the shifts we are looking for may be being constrained by some of the facilitators' beliefs and ways of working and that we need to search for new ways of communicating. From a project perspective, we need to ask:

- How can we communicate to facilitators that what we did with the first cohort was not good enough to make some of the improvements that we need to make?
- How do we examine our beliefs and support facilitators to examine theirs, as these may be the constraints for the shifts we are looking for?

These questions will be explored in the "Project response" sections of the analysis.

Engaging others

The project needs to engage others in this process of inquiry and take part in evidence-based conversations.

We want to engage others in exploring the complexity of school improvement and of achieving sustainable improvement in student literacy outcomes. This report presents as much information as possible about the improvements in student achievement that the project has achieved, as well as the areas of concern. This information is underpinned with descriptions of the limitations of the tools and their uses. We hope that by doing this, the report format will enable others to engage in the learning. This could be by critiquing the project and suggesting other possible evidence-based project

[4] Education Review Office (2003). *Evaluation Indicators for Education Reviews in Schools*. page 41.
[5] Waters, T., Marzano, R.J., and McNulty, B. (2003). "Balanced Leadership: What 30 Years of research tells us about the effect of leadership on student achievement". Unpublished paper. McREL.

responses or areas of concern, or it could be by exploring ways of transferring some of the learning to other projects.

Identifying shifts in the outcomes

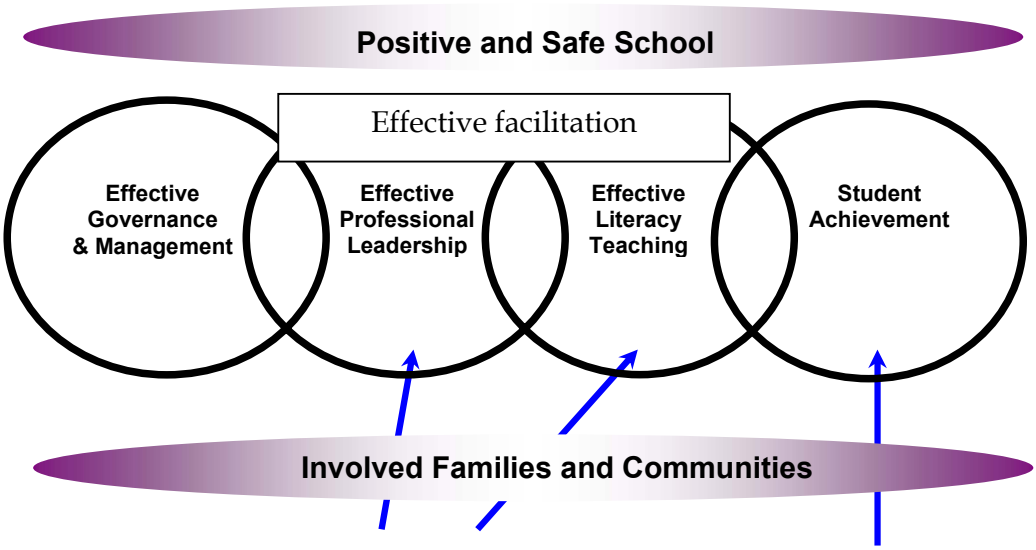
The project needs to identify shifts in the outcomes for the project as a whole.

The project is accountable. This means "providing transparent and informative statements of account to others … active self monitoring … and a sense of responsibility for the quality of work" (National College for School Leadership, 2004)[6] across the project.

[6] National College for School Leadership (2004). "What makes a network a learning network?" Downloaded from www.ncsl.org.uk on 10 December 2004.

Figure 2: Linking the project outcomes

Chain of Quality – Evaluation Indicators



Process Indicators

Proxies for the desirable outcomes

- Evidence of improved teacher literacy content knowledge
- Evidence of transfer of understanding of literacy pedagogy to practice
- Evidence of professional learning communities

Outcome Indicators

Direct outcomes

- Evidence of improved student literacy achievement

A description of the limitations of the needs analysis tools

The two tools, STAR for reading comprehension and asTTle for writing, were chosen because they provide teachers with professional learning about literacy and about the students. They also provide direction about the response teachers should make to their students' needs. Both tools encourage teachers to ask the question, "How effective is my teaching in improving the learning on achievement of all students?"

Many facilitators were unfamiliar with one or both tools. They were unfamiliar with the theory behind what each tool was testing and what inferences could be made from the data, as well as the administration, marking, and moderation associated with each test. During 2004 the facilitators were provided with ongoing professional learning in the use of asTTle at regional meetings and at the four national hui. There were two major sessions on STAR at the first two hui in 2004.

The project has had a focus on schools developing data that they can trust so that the interpretations about the effectiveness of teaching can be trusted. Because of this, we have assumed that the data entry has been checked at the school level.

Reading comprehension: STAR

Limitations to the tool

Many schools with a reading focus appeared to move through phase 1 quite quickly. The information from the analysis did not appear to provide facilitators and schools with a rich focus for building literacy content knowledge.

The tool appears to have a ceiling effect for year 6 students and those students at stanines 7–9. This is explored further in the actual analysis.

Limitations in the use of the tool

Once guidelines for administering STAR had been established with the teachers, the administration, marking, and recording of scores seemed relatively straightforward. Some facilitators spoke anecdotally of schools using the wrong set of national data when identifying students with critical scores or determining stanines. In addition, some schools appeared not to have systems for checking teachers' copying of scores or the addition of scores to determine a total. We did sample-check the addition of subtest scores for the total score and then for the correct stanine. There were a few errors in teachers' addition of the subtest scores. These were rectified before conducting the analysis. There were no mistakes in the stanine allocations.

Our experience suggests that STAR has room for error in:

- the copying of scores to a class/year-group table;
- the addition of subtest scores;

- the application of the national data to the scores for the initial analysis (such as critical score range and stanines).

Many schools and facilitators had very rudimentary statistical knowledge and were unable to explore many of the patterns and anomalies within the data, especially at Time 1. Time was allocated at a national seminar for the facilitators to develop their skills in using Excel and to think about the sorts of analysis schools would be interested in.

Many schools and facilitators used the subtest score information to identify the target students and to group other students. They needed a stronger knowledge base to use the information to ask about the effectiveness of their practice and what they need to know about reading comprehension.

Writing: asTTle

Limitations in the tool

The administration, marking, and scoring of asTTle was complex because it relied on the markers having a very sound knowledge of writing features as described in the asTTle indicators. Each facilitator devised a system for teacher learning as the schools marked and moderated the students' scripts. At this point, there was no guidance from asTTle about moderation.

We do not know whether schools sampled the accuracy of the data they entered into asTTle.

Ten percent of the students' results were sent as a sample to Learning Media as an asTTle file. It was expected that the facilitators had moderated all or a sample of these scripts. The accuracy of the marking did rely on the facilitators' knowledge. In the asTTle report[7] on marking, the researchers found that on average teachers were within the acceptable range of the assessment leader's score 75% of the time.

Therefore, it appears that asTTle has room for error in:

- marking student scripts;
- copying the scores into the asTTle database.

Limitations in the use of the tool

The asTTle marking study found that teachers needed support in the areas of grammar and language resources. This is also the anecdotal evidence from regional meetings. Many facilitators did not find the example scripts within asTTle very useful in supporting both their and the teachers' understanding of the marking criteria. We cannot guarantee the accuracy of the initial marking and subtest data entry, as schools carried out the entry. Again, because there has been a focus on developing trustworthy data, we assume that the processes used made the data as accurate as possible and that Time 3 data is probably more accurate than Time 1 data.

[7] asTTle Technical Report 26: Accuracy in the scoring of writing

For some teachers and schools, the technical issues with asTTle outweighed the benefits.

All schools used asTTle version 3 at Time 1, most schools used asTTle version 4 at Time 3, but some used the earlier version at Time 3. There were 902 students who were assessed using version 3 at Time 1 and version 4 at Time 3 and 162 students using version 3 at both points in time. The major difference between these two versions is the level 4 indicators in the marking schedule. A piece of student writing at level 4 + would have ended up with a higher asTTle: Writing score (aWs) if they were marked using the version 4 indicators than if the version 3 indicators were used. When looking at the data on a year-by-year basis, all but one of the year groups did better when assessed on the version 4 indicators than those who were assessed using the version 3 indicators. The exception was the year 6 group. Fifty-nine year 6 students who were assessed both times using the version 3 indicators had a much higher mean asTTle score at Time 3 than those who were assessed using version 4. This needs to be explored further, especially for scores greater than 638, to determine its impact on the analysis.

Schools chose their writing test for a variety of reasons. Most schools chose a test with a focus on the writing area they were going to teach next for Time 1 but may have had a both a formative and a summative purpose for Time 3 as the test was carried out late in term 3 or early in term 4, 2005.

A description of who is represented in each data set

Data has been gathered for this analysis from eighty-six schools at two points in time over the two years from February 2004 to December 2005. Time 1 was term 1, 2004 and Time 3 was term 4, 2005. Data has only been used for students who were at the schools at both Time 1 and Time 3. This means there may be quite a large group of students who were not in the schools at either Time 1 or Time 3 (that is, they were there at only one of the points of time) that we do not know anything about.

It was expected that schools would assess all students except those that are ORS-funded or very recent non-English speaking immigrants. We did not ask schools to identify the students they did not assess so do not know who was or was not assessed.

Reading focus schools

Forty of the forty-two reading focus schools had data at both Time 1 and Time 3. One school lost its Time 1 data and another school did not collect data at Time 3. This school had been identified as one that was at risk in June 2005. The forty schools had a total of 3787 students from years 3 to 8. The following three tables show the demographic information for the schools.

Table 5: Year groups

Year level at Time 1	*Number of students*	*% of students*
3	498	13.2
4	511	13.5
5	489	12.9
6	167	4.4
7	2122	56.0
All	**3787**	**100**

The students in years 3 to 5 are at full primary and contributing schools. The year 6 students are at full primary schools only. These students are under-represented in the sample, even though 44% of the schools were full primary. The year 7 students are over-represented in the sample. These students come from full primary, intermediate, and years 7 to 15 schools.

Table 6: Gender

Gender	*Number of students*	*% of students*
Male	1974	52.1
Female	1813	47.9
All	**3787**	**100**

Table 7: Ethnicity

Ethnicity	*Number of students*	*% of students*
NZ European/Pakeha	2170	57.3
NZ Māori	890	23.5
Pasifika	356	9.4
Asian	284	7.5
Other	40	1.1
Other European	47	1.2
All	**3787**	**100**

Writing focus schools

Forty-five of the forty-nine schools had data at both Time 1 and Time 3. This represents a total of 1063 students. One of the schools without data at Time 3 was a school identified as at risk in June 2005. Two schools lost their Time 1 data. The fourth school is a residential school and has students for only one year. It has data from both years but for two quite different cohorts of students.

The writing data is from a 10% random sample of students or fifty students – whichever is the greatest – from each school. When comparing student numbers with the reading data, this percentage could be closer to 30% of the students. Schools with small student numbers (fewer than fifty) will be over-represented in this sample as data from all the students will be in the national picture. asTTle has been designed for years 4–8 students but, as shown below, some schools use asTTle to diagnose year 3 students' writing strengths and needs.

The following three tables show the sample's demographic information.

Table 8: Year groups

Year level at Time 1	*Number of students*	*% of students*
3	6	0.6
4	347	32.6
5	332	31.2
6	134	12.7
7	244	22.9
All	**1063**	**100**

As with the reading data, there is a smaller group of year 6 students compared with the other year groups. Year 7 is not over-represented as it was with the reading comprehension data.

Table 9: Gender

Gender	*Number of students*	*% of students*
Male	525	49.4
Female	538	50.6
All	**1063**	**100**

Table 10: Ethnicity

Ethnicity	*Number of students*	*% of students*
NZ European/Pakeha	723	68.0
NZ Māori	192	18.1
Pasifika	70	6.6
Asian	47	4.4
Other	22	2.1
Other European	8	0.8
Unknown	1	0.1
All	**1063**	**100**

There is a higher percentage of NZ European/Pakeha students in the writing sample than in the reading comprehension data. The percentage of students in the NZ Māori,

Pasifika, and Asian ethnic groups is less than it was for the reading comprehension data.

An evidence-based evaluation of the overall effectiveness of the project

The focus of this first analysis and interpretation of the student data is to identify the shifts in achievement for the cohort as a whole and then to identify whether or not particular subgroups within the cohort have made similar shifts.

The project has a focus on the "tail", as identified in national and international testing. Teachers have used students in the "tail" as the "touchstone" for their effectiveness as they trialled deliberate acts of teaching throughout the project. Therefore, the analyses of this particular subgroup's shifts in achievement are important when reflecting on the effectiveness of the project. The Time 1 and Time 3 scores for the students in the lowest 20% of students for writing and lowest three stanines for reading at Time 1 have been compared to monitor the shifts each of these groups have made.

Reading focus schools

What is the mean stanine shift for each year group?

After two years in the project, the mean stanine for the cohort as a whole was 5.64, with a small range of 0.25 across years (from 5.57–5.82). The Time 1 mean was 5.08, with a much larger range of 0.85 (from 4.89–5.74). The shift in mean stanine for the cohort as a whole was 0.56 across year groups. Years 4 and 7 shared the largest shift of 0.7 of a stanine, and year 6 students had the smallest mean shift of 0.08. The year 6 students had the highest mean stanine at Time 1 and Time 3.

Effect sizes cannot be calculated for stanines so they are given as a measure based on the raw scores. The effect size, using these raw STAR scores, was determined for each year group. The mean effect size over the two years was 0.87. Year 4 had the largest effect size of 1.12, and year 3 had the smallest of 0.54.

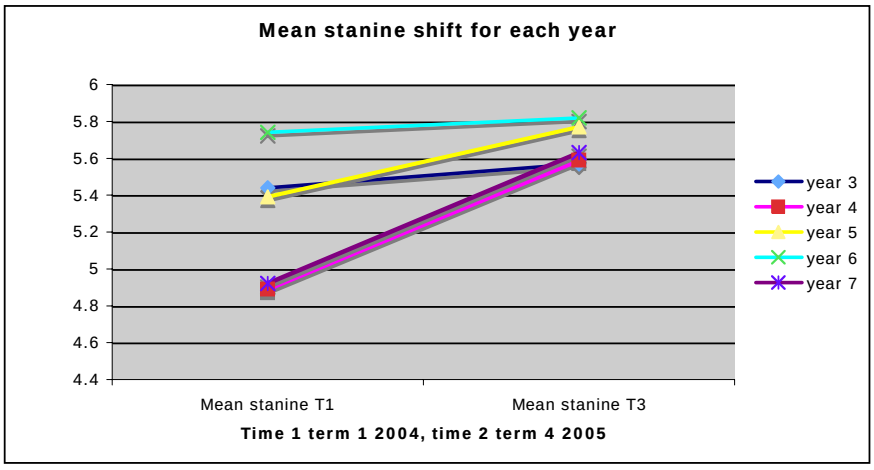
This interpretation is based on the descriptive statistics analysis shown in Table 11 and represented graphically in Figure 3.

Table 11: Mean stanine at Time 1 and Time 3 for each year

Year level at Time 1	*N*	*Mean stanine T1*	*Mean stanine T3*	*Difference*	*Effect Size*[a]
3	498	5.44	5.57	0.13	0.54
4	511	4.89	5.59	0.70	1.12
5	489	5.39	5.77	0.38	0.95
6	167	5.74	5.82	0.08	-[b]
7	2122	4.92	5.63	0.71	0.89
All	**3787**	**5.08**	**5.64**	**0.56**	**0.87**

[a] Effect size is measured using the raw scores rather than the stanines.
[b] No effect size is given for year 6 as we cannot compare the raw scores for STAR Test 4–6 (taken in year 6) and STAR Test 7–9 (taken in year 7).

Figure 3: Mean stanine shift for each year



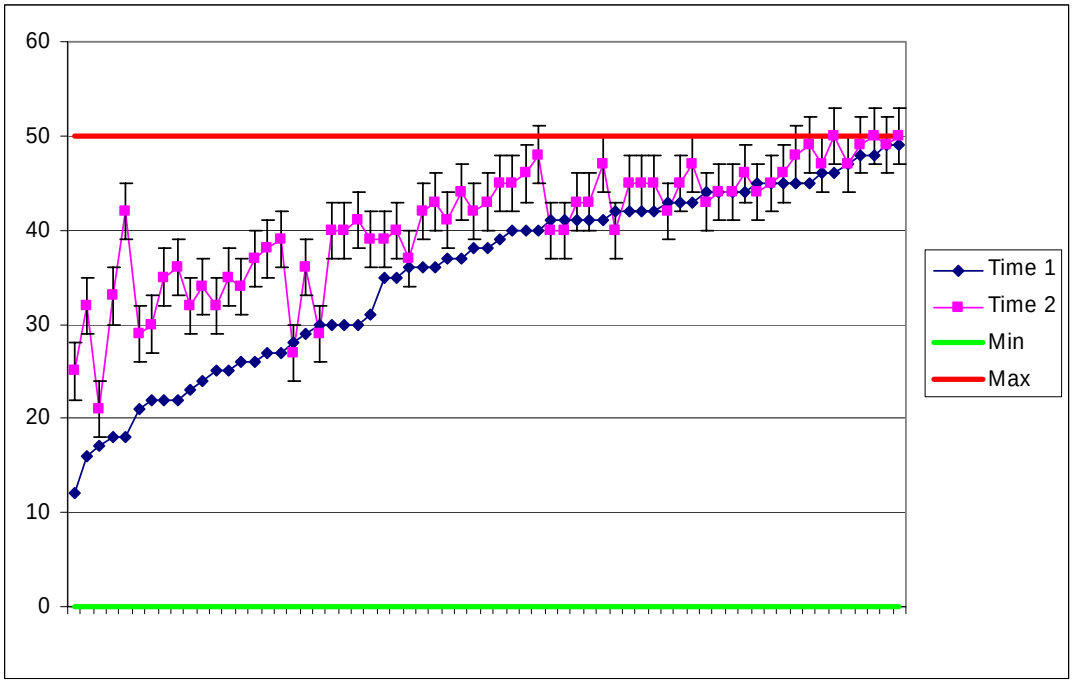
0

As discussed above and shown in Figure 3, the year 6 students had the highest mean stanine scores at both Time 1 and Time 3 and yet had the smallest gain over this time. This small gain can be understood if considered in terms of a "ceiling effect". A ceiling effect occurs when an upper limit on a test results in a negatively skewed distribution. A negatively skewed distribution is helpful for identifying people who are at risk, distinguishing those individuals with a lower true score than that of the general population (Cronbach, 1990)[8] – something STAR was designed to do. However, ceiling effects are problematic in distinguishing those at the upper end and in accurately measuring improvement. If someone received the maximum value of a test at Time 1 testing, then they will of course show no improvement whether or not their "true

[8] Cronbach Lee J. (1990). *Essentials of Psychological Testing*. Harper & Row: New York, page 212.

ability" has improved. For the year 6 students, the majority of scores are at the higher end of the test scale at or near the maximum value. Therefore, there is very little room for improvement with STAR. See Figure 4. This ceiling effect does not seem to be as apparent with the other year groups.

Figure 4: Year 6 STAR data T1 and T3



What is the mean stanine shift for students in each of the lowest 3 stanines at Time 1 for each year group?

The mean stanine shift was then explored for the students in the first three stanines for each year group at Time 1. There were 22.32% of all students in these three stanines. The year 6 students had the largest shift over the two years (1.5 stanines). This is congruent with the notion of a ceiling effect operating for the students in the upper stanines whereas, arguably, the students in the lowest stanines had room to grow. Mean shifts ranged from 1.5 to 0.68, with the students in the lowest three stanines of all year groups making a much larger shift than their year group as a whole. This is shown in Table 13 and Figure 5.

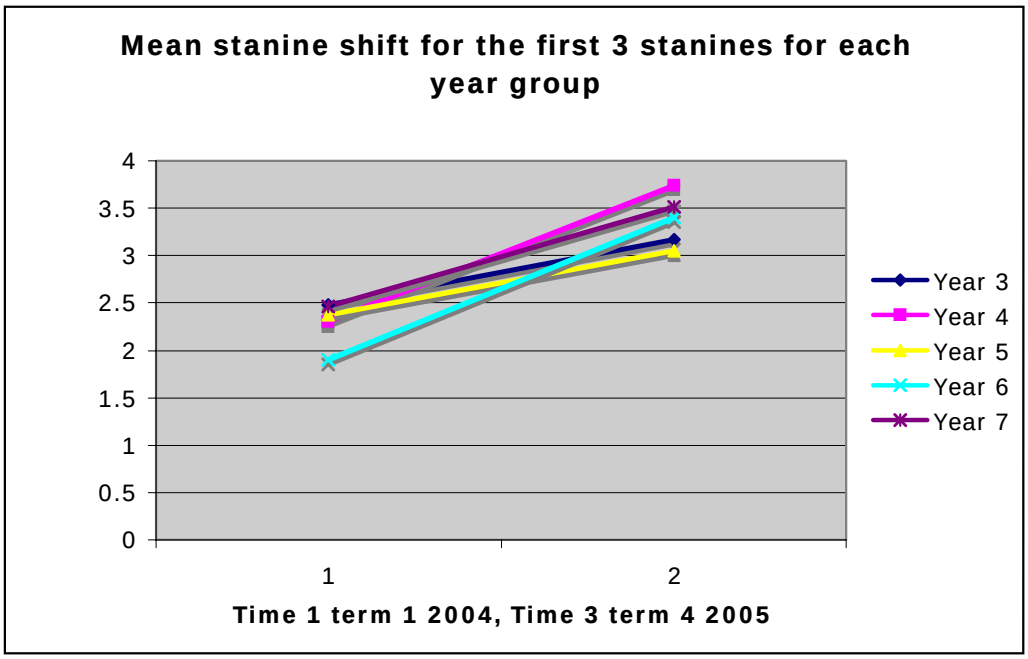
The effect size, based on the raw scores, ranged from 1.75 to 2.05 with a mean effect size for students in stanine 1–3 at Time 1 for all year groups of 1.97. The effect size based on the raw scores for all students in each year group was 0.87.

Table 12: Mean stanine at Time 1 and Time 3 for each year for students in stanines 1–3 at Time 1

Year level at Time 1	N	Mean stanine T1	Mean stanine T3	Difference	Effect size[a]
3	94	2.48	3.17	0.69	1.75
4	117	2.30	3.74	1.44	2.31
5	82	2.37	3.05	0.68	2.05
6	20	1.90	3.40	1.50	-[b]
7	532	2.46	3.51	1.06	1.78
All Stanines 1 - 3	**845**	**2.42**	**3.46**	**1.04**	**1.97**

[a] Effect size is measured using the raw scores not the stanines.
[b] No effect size is given for year 6, as we cannot compare the raw scores for STAR Test 4–6 (taken in year 6) and STAR Test 7–9 (taken in year 7).

Figure 5: Mean stanine at Time 1 and Time 3 for each year for students in stanines 1–3 at Time 1



STAR is a standardised test that uses "the scores of a large representative sample of students to establish stanine norms for each class level" NZCER (2001) STAR[9] Supplementary tests of achievement in reading years 4-9. Wellington. NZCER. As explained earlier with stanines, it is assumed that if a student were to maintain their current rate of progress (with general teaching and student development) there would be no shift in a group of students' individual stanine scores over time and, therefore, no improvement in their mean stanine scores. Any shift seen in a year group's mean stanines show an altered trajectory and, in this case, the change is hypothesised to be attributed to increased teaching effectiveness. Further analysis is needed to know

[9] NZCER (2001). *STAR Supplementary tests of achievement in reading years 4 - 9*. Wellington. NZCER.

whether the shift seen in the mean stanines for each year group as described above can be associated with the increased effectiveness of teaching or can be explained by the statistical phenomena of regression to the mean.

What is the shift from each stanine?

The data was analysed in two different ways. Initially, the number of students in each stanine at Time 1 and Time 3 were compared. This is explored in the first section below and shown in Table 14 and Figure 6. The project was particularly interested in the shifts in achievement for the students in the stanine 1–3 band at Time 1 and whether their collective shift was different to the cohort as a whole. This analysis is explored in the second section below and shown in Tables 15–17 and Figure 7.

1. Number and percentage of students in each stanine at Time 1 and Time 3

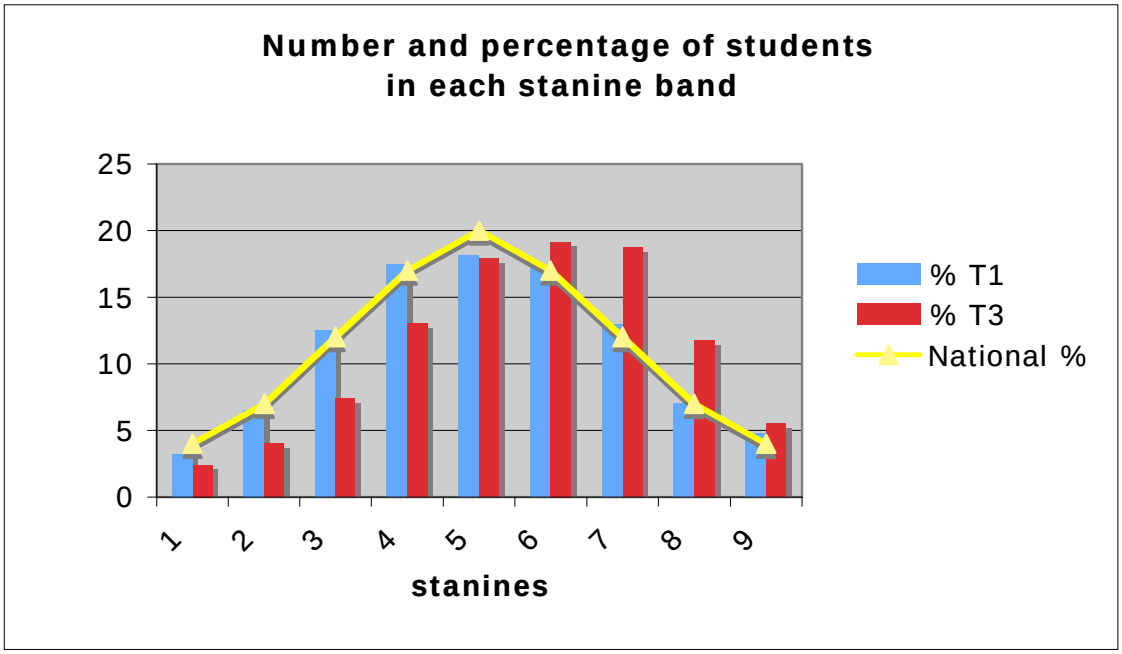
There has been a shift in the number of students and percentage of the total within each stanine. This shift is a negative one from stanines 1 to 5 where there is a loss of students and a positive one from stanines 6 to 9. Stanines 3, 4, 7, and 8 have had the largest gains in the percentage of students from Time 1 to Time 3.

Table 13 and Figure 6 show the analysis on which this interpretation is based.

Table 13: Number and percentage of students in each stanine band at Time 1 and Time 3

Stanine	*N T1*	*% T1*	*N T3*	*% T3*	*Difference in %*	*National %*
1	123	3.25	92	2.43	-0.82	4
2	248	6.55	153	4.04	-2.51	7
3	474	12.52	281	7.42	-5.10	12
4	661	17.45	494	13.04	-4.41	17
5	686	18.11	678	17.90	-0.21	20
6	658	17.38	725	19.14	1.76	17
7	490	12.94	709	18.72	5.78	12
8	266	7.02	446	11.78	4.76	7
9	181	4.78	209	5.52	0.74	4
All	**3787**	**100**	**3787**	**100**		

Figure 6: Number and percentage of students in each stanine band at Time 1 and Time 3



2. *The shift in achievement from Time 1 to Time 3*

We were particularly interested in the shifts individual students made from Time 1 to Time 3. Some students had a negative shift, others no shift, while others shifted 1 or 2+ stanines. Table 7 shows that more students in the first three stanines at Time 1 had a shift of 2+ than in the other two stanine bands. The group in stanines 4–6 at Time 1 had more students with a shift of one stanine than the other two bands, and the group in stanines 7–9 at Time 1 had more students with a negative shift or no shift than the other two bands.

For individual students there needs to be a stanine shift of at least two stanines for "there to be good reason to believe that the difference is a real one, and not due to chance factors in the testing process" (NZCER, 2001, page 21)[10]. It can be shown that a stanine shift of at least two is equivalent to an effect size of at least one given the formulae underlying the calculation of stanine scores. Of the students most at risk of continuously underachieving, 40% of stanine 1 students made a significant shift that can be attributed to effective teaching and 34% of stanine 2 students made a significant shift. The percentage of students with a 2+ stanine shift decreased as the stanines increased.

Table 14 also shows the lack of shift for students at stanines 7 to 9 at Time 1. Approximately 43% of these students made no shift over the two years and 55% of students at stanine 9 Time 1 made a negative shift. Earlier we explored the ceiling

[10] NZCER (2001). *STAR Supplementary tests of achievement in reading years 4 - 9.* Wellington. NZCER.

effect for year 6 students. Further analysis would allow us to know whether this is the reason for the lack of shift shown by these students or whether the project has had a negative impact on learning for these students.

Table 14: **Percentage of students for each shift in stanine**

Stanine at T1	*Negative shift*	*No shift*	*1 stanine shift*	*2+ stanine shift*
1	0.00	36.59	24.39	39.02
2	12.10	26.61	27.02	34.27
3	10.97	26.79	33.12	29.11
4	9.38	29.95	36.31	24.36
5	11.81	29.59	35.42	23.18
6	15.81	34.50	33.89	15.81
7	23.06	42.65	26.53	7.76
8	37.22	43.23	19.55	0.00
9	55.25	44.75	0.00	0.00
All	**16.93**	**33.56**	**30.16**	**19.36**

Table 15 shows in more detail the shifts from Time 1 to Time 3 as a percentage of Time 1. Although the project can celebrate the fact that a high proportion of students in the stanine 1–3 band had a shift of 2+ stanines, there is a concern that 84.5% of the students at stanine 1 at Time 1 are still within this band of stanines at Time 3. We found 65.7% of students at stanine 2 at Time 1 are still within this band of stanines at Time 3, and 38% of the Time 1 stanine 3 students are still within this band.

Table15:Stanine movement from Time 1 to Time 3 (percentage of Time 1 stanine)

% students at T1 stanine		*Stanine T3*								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Stanine T1	**1**	**36.59**	24.39	23.58	11.38	1.63	0.00	1.63	0.81	0.00
	2	12.10	**26.61**	27.02	17.74	10.89	2.82	1.21	1.61	0.00
	3	2.74	8.23	**26.79**	33.12	18.78	5.70	2.32	1.90	0.42
	4	0.00	2.27	7.11	**29.95**	36.31	16.64	5.75	1.36	0.61
	5	0.29	0.44	1.02	10.06	**29.59**	35.42	17.78	4.23	1.17
	6	0.15	0.00	0.30	1.67	13.68	**34.50**	33.89	12.16	3.65
	7	0.00	0.00	0.20	0.00	5.10	17.76	**42.65**	26.53	7.76
	8	0.38	0.00	0.00	0.38	0.75	7.14	28.57	**43.23**	19.55
	9	0.00	0.00	0.55	0.00	0.00	2.76	13.81	38.12	**44.75**

When we unpack stanine 1 into year groups, it appears that 3% of the students in each year group were in stanine 1 at Time 1, with the exception of years 4 and 6 where the figure was 5%.

Table 16: shows the percentage of stanine 1 students for each year group moving from Time 1 to Time 3: All of the year 3 stanine 1 students remain in the stanine 1–3 band, 69% of years 4 and 5 stanine 1 students remain in the stanine 1–3 band, and approximately 88% of years 6 and 7 stanine 1 students remain in the stanine 1–3 band.

Table 16: Stanine 1 movement from Time 1 to Time 3 (percentage of Time 1, stanine 1)

% of students at T1 stanine		*Stanine T3*								
		1	2	3	4	5	6	7	8	9
Stanine 1 T1 Year	3	23.08	46.15	30.77	0.00	0.00	0.00	0.00	0.00	0.00
	4	26.09	21.74	21.74	21.74	8.70	0.00	0.00	0.00	0.00
	5	30.77	23.08	15.38	30.77	0.00	0.00	0.00	0.00	0.00
	6	25.00	12.50	50.00	12.50	0.00	0.00	0.00	0.00	0.00
	7	45.45	22.73	21.21	6.06	0.00	0.00	3.03	1.52	0.00

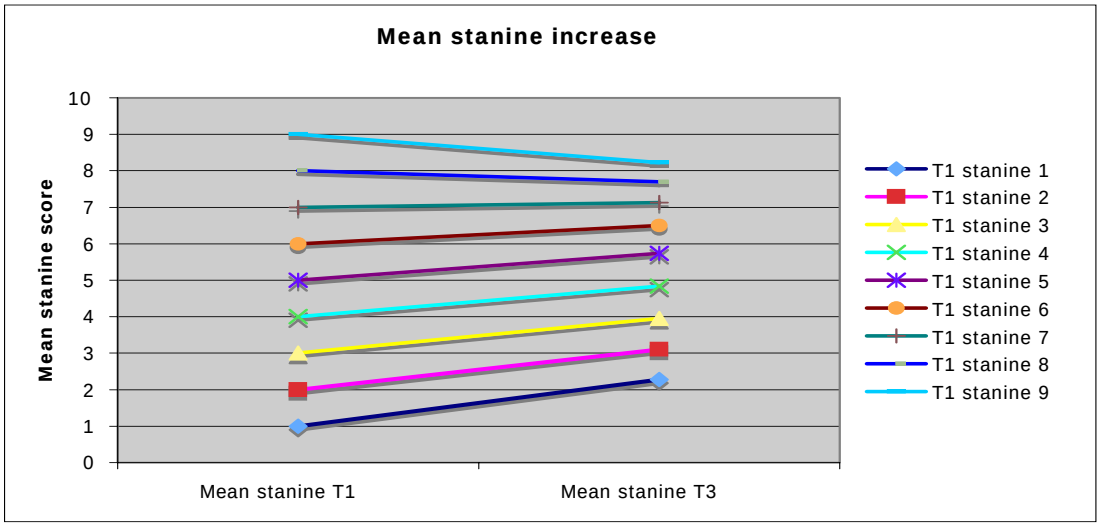
The students in the first three stanines made a much larger shift in mean stanine than the cohort as a whole. At a later date we will analyse why the students at stanines 8 and 9 at Time 1 made no shift or a negative shift. We will test the assumption that STAR has a ceiling effect for students in these two stanines.

Table 17 and Figure 7 show the shift in achievement for each stanine group. The reading comprehension schools within the project have made a significant shift in student achievement for the cohort as a whole but, more importantly, they appear to have accelerated the rate of progress for students in the lowest three stanines. This rate of progress is offset by what appears to be a lack of progress for the students in the highest three stanines. Again, this could be due to a number of factors such as the measurement error in STAR, or STAR having a ceiling effect, or there is a regression to the mean. The shifts by both stanine groups (the highest and the lowest) could be explained by the regression to the mean phenomena. This requires further analysis.

Table 17: Mean stanine increase from each stanine

Stanine at T1	*N T1*	*Mean stanine T1*	*Mean stanine T3*	*Difference*
1	123	1.00	2.28	1.28
2	248	2.00	3.10	1.10
3	474	3.00	3.95	0.95
4	661	4.00	4.84	0.84
5	686	5.00	5.74	0.74
6	658	6.00	6.50	0.50
7	490	7.00	7.13	0.13
8	266	8.00	7.70	-0.30
9	181	9.00	8.23	-0.77
All	**3787**	**5.08**	**5.64**	**0.56**

Figure 7: Mean stanine increase



What shifts do we see when the data is analysed through the lens of gender?

The mean stanine shift for boys is 0.6, which is larger than the cohort as a whole (0.56). The mean stanine for boys at Time 3 is equal to the Time 1 mean stanine for girls. The evidence for this interpretation is shown in Table 18.

Table 18: Mean stanine at Time 1 and Time 3

Gender	*N*	*Mean stanine T1*	*Mean stanine T3*	*Difference*
Male	1974	4.79	5.39	0.60
Female	1813	5.40	5.92	0.52
All	**3787**	**5.08**	**5.64**	**0.56**

Tables 19 and 20 show that boys are over-represented in the first four stanines at Time 1 and the first five stanines at Time 3.

Table 19: Percentage of males and females within stanines at Time 1

Stanine at T1	*N Males*	*% Males*	*N Females*	*% Females*	*Total*
1	88	71.54	35	28.46	123
2	175	70.56	73	29.44	248
3	292	61.60	182	38.40	474
4	368	55.67	293	44.33	661
5	330	48.10	356	51.90	686
6	317	48.18	341	51.82	658
7	202	41.22	288	58.78	490
8	113	42.48	153	57.52	266
9	89	49.17	92	50.83	181
All	**1974**	**52.13**	**1813**	**47.87**	**3787**

Table 20: Percentage of males and females within stanines at Time 3

Stanine at T3	*N Males*	*% Males*	*N Females*	*% Females*	*Total*
1	70	76.09	22	23.91	92
2	100	65.36	53	34.64	153
3	174	61.92	107	38.08	281
4	297	60.12	197	39.88	494
5	368	54.28	310	45.72	678
6	330	45.52	395	54.48	725
7	337	47.53	372	52.47	709
8	203	45.52	243	54.48	446
9	95	45.45	114	54.55	209
All	**1974**	**52.13**	**1813**	**47.87**	**3787**

Figures 8 and 9 show the percentage of boys and girls at each stanine for Time 1 and Time 3. For both groups, there is a shift to a more positively skewed distribution shown in the graphs, with stanines 3, 7, and 8 having the largest shift for boys and stanines 4, 7, and 8 having the largest shift for girls.

Figure 8: Percentage of boys at each stanine, Time 1 and Time 3

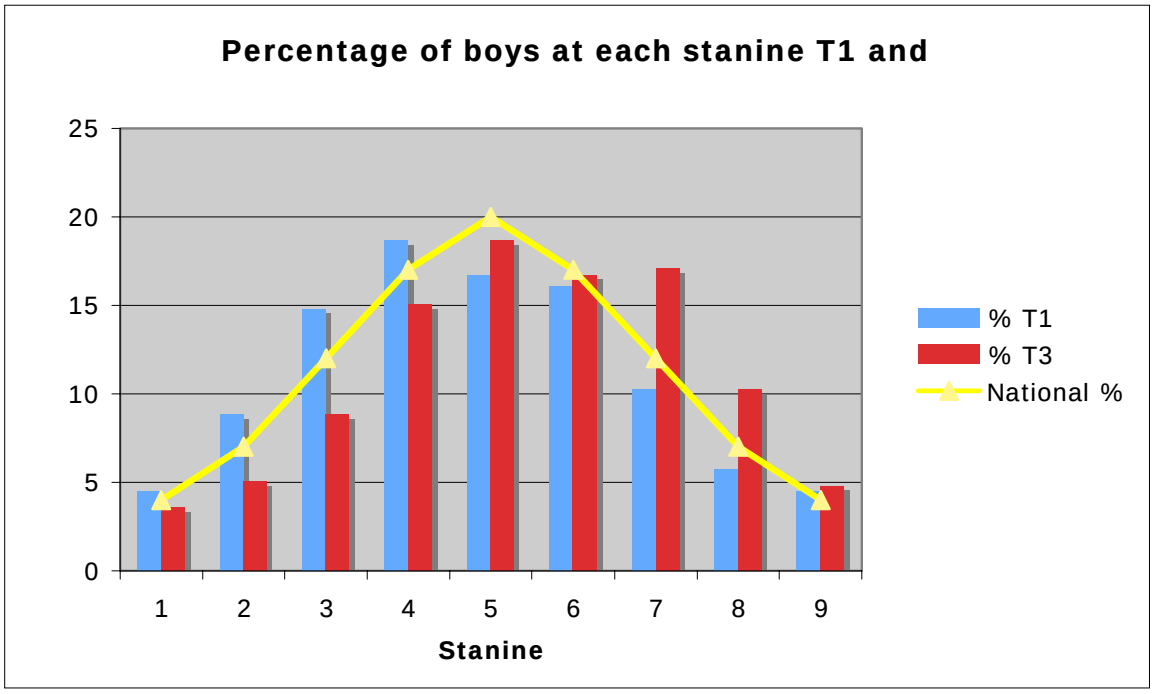
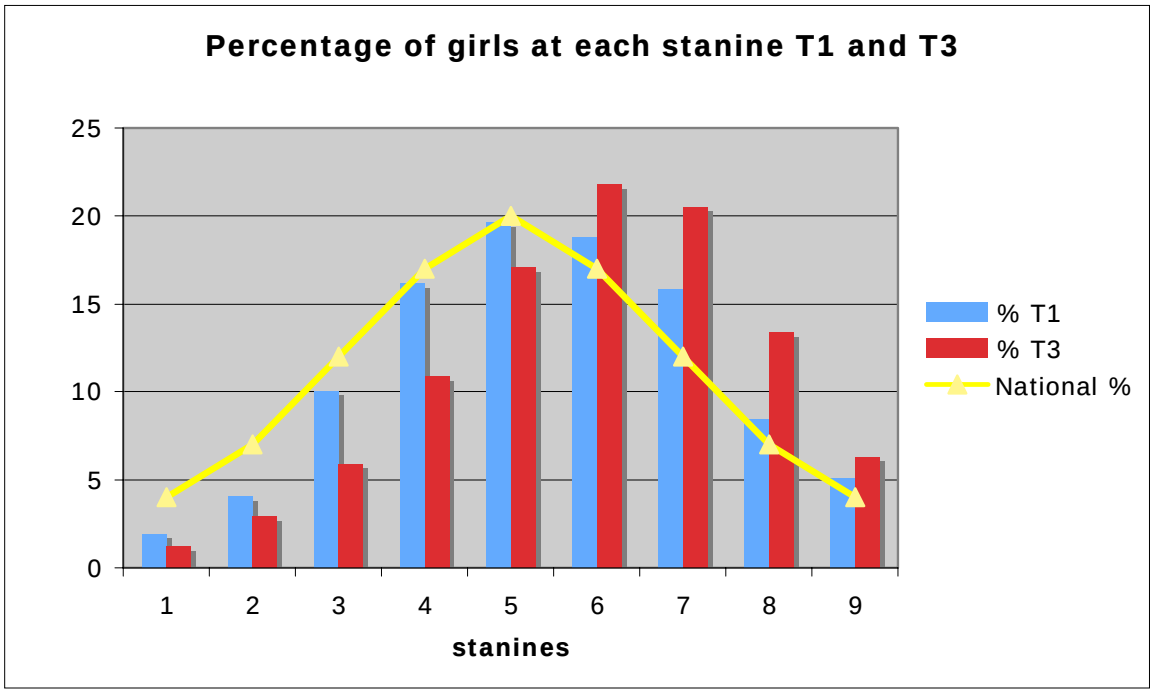


Figure 9: Percentage of girls at each stanine, Time 1 and Time 3



Forty-five percent of the girls at stanine 1 at Time 1 had a 2+ stanine shift, compared with only 36% of the boys. Therefore, 74% of the girls at stanine 1 at Time 1 are still within the first three stanines at Time 3, compared with 88.6% of the boys. The percentage of boys not shifting is 4% less than for the cohort as a whole. See Tables 21 and 22 for the descriptive statistics that this is based on.

Table 21: Stanine movement from Time 1 to Time 3: Boys (percentage of Time 1 stanine)

Boys

% of students at T1 stanine		*Stanine T3*								
		1	2	3	4	5	6	7	8	9
Stanine										
T1	1	42.05	21.59	25.00	6.82	1.14	0.00	2.27	1.14	0.00
	2	12.00	26.86	25.14	19.43	10.86	2.29	1.14	2.29	0.00
	3	3.42	8.22	24.32	34.25	19.86	4.11	3.42	1.71	0.68
	4	0.00	2.72	7.88	29.62	35.87	16.85	5.71	0.82	0.54
	5	0.30	0.00	1.21	12.12	29.70	32.42	20.61	3.33	0.30
	6	0.32	0.00	0.63	2.21	15.46	31.55	34.07	13.25	2.52
	7	0.00	0.00	0.50	0.00	4.46	16.34	40.10	29.21	9.41
	8	0.00	0.00	0.00	0.88	1.77	7.96	29.20	42.48	17.70
	9	0.00	0.00	1.12	0.00	0.00	3.37	13.48	33.71	48.31

Table 22: Stanine movement from Time 1 to Time 3: Girls (percentage of Time 1 stanine)

Girls

% of students at T1 stanine		*Stanine T3*								
		1	2	3	4	5	6	7	8	9
Stanine										
T1	1	22.86	31.43	20.00	22.86	2.86	0.00	0.00	0.00	0.00
	2	12.33	26.03	31.51	13.70	10.96	4.11	1.37	0.00	0.00
	3	1.65	8.24	30.77	31.32	17.03	8.24	0.55	2.20	0.00
	4	0.00	1.71	6.14	30.38	36.86	16.38	5.80	2.05	0.68
	5	0.28	0.84	0.84	8.15	29.49	38.20	15.17	5.06	1.97
	6	0.00	0.00	0.00	1.17	12.02	37.24	33.72	11.14	4.69
	7	0.00	0.00	0.00	0.00	5.56	18.75	44.44	24.65	6.60
	8	0.65	0.00	0.00	0.00	0.00	6.54	28.10	43.79	20.92
	9	0.00	0.00	0.00	0.00	0.00	2.17	14.13	42.39	41.30

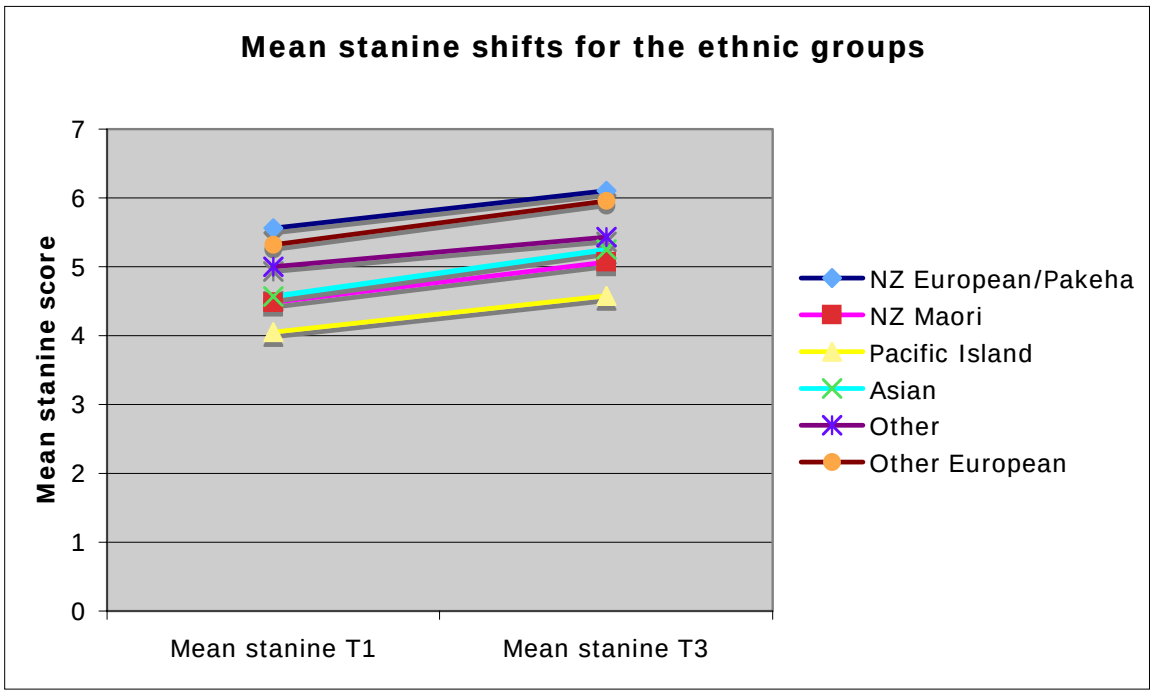
What shifts do we see when the data is explored through the lens of ethnicity?

Table 23 and Figure 10 show that NZ Māori, Asian, and Other European students had a bigger mean stanine shift than the cohort as a whole. These shifts bring the NZ Māori mean stanine score at Time 3 close to the Time 1 mean stanine score for the whole cohort. The question is whether this is good enough. The smaller shift by the NZ European/Pakeha may be explained by the STAR ceiling effect, as the group had the highest mean stanine score at both times. What is of concern is that Pasifika students have a mean stanine shift that is less than the cohort as a whole and have the lowest mean stanine at both Time 1 and Time 3.

Table 23: Mean stanine at Time 1 and Time 3

Ethnicity	*N*	*Mean stanine T1*	*Mean stanine T3*	*Difference*
NZ European/Pakeha	2170	5.56	6.10	0.54
NZ Māori	890	4.48	5.07	0.59
Pasifika	356	4.05	4.58	0.53
Asian	284	4.57	5.25	0.68
Other	40	5.00	5.43	0.43
Other European	47	5.32	5.96	0.64
All	**3787**	**5.08**	**5.64**	**0.56**

Figure 10: Mean stanine shifts for the ethnic groups



The following four figures show the percentage of each ethnicity in each stanine at Time 1 and Time 3. The biggest shifts for each ethnicity were: NZ European stanines 4, 7, and 8; Māori stanines 3, 7, and 8; Pasifika stanines 3, 5, and 6; and Asian stanines 2, 6, and 7.

Figure 11: Percentage of students in each stanine at Time 1 and Time 3, NZ European

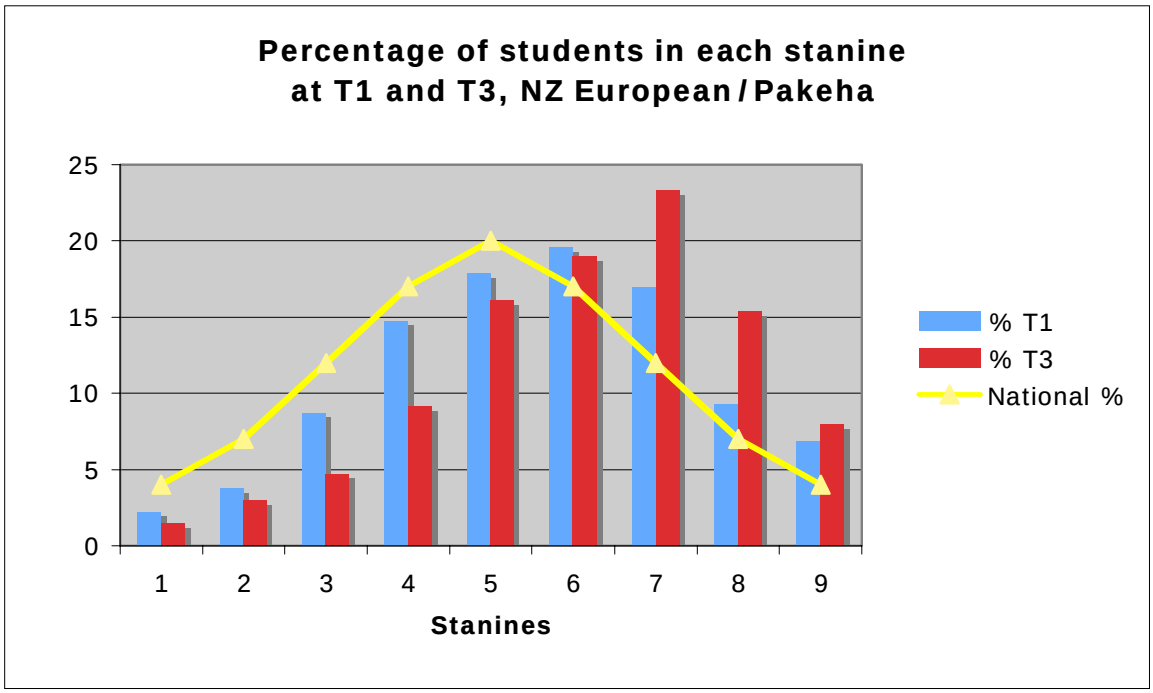


Figure 12: Percentage of students in each stanine at Time 1 and Time 3, Māori

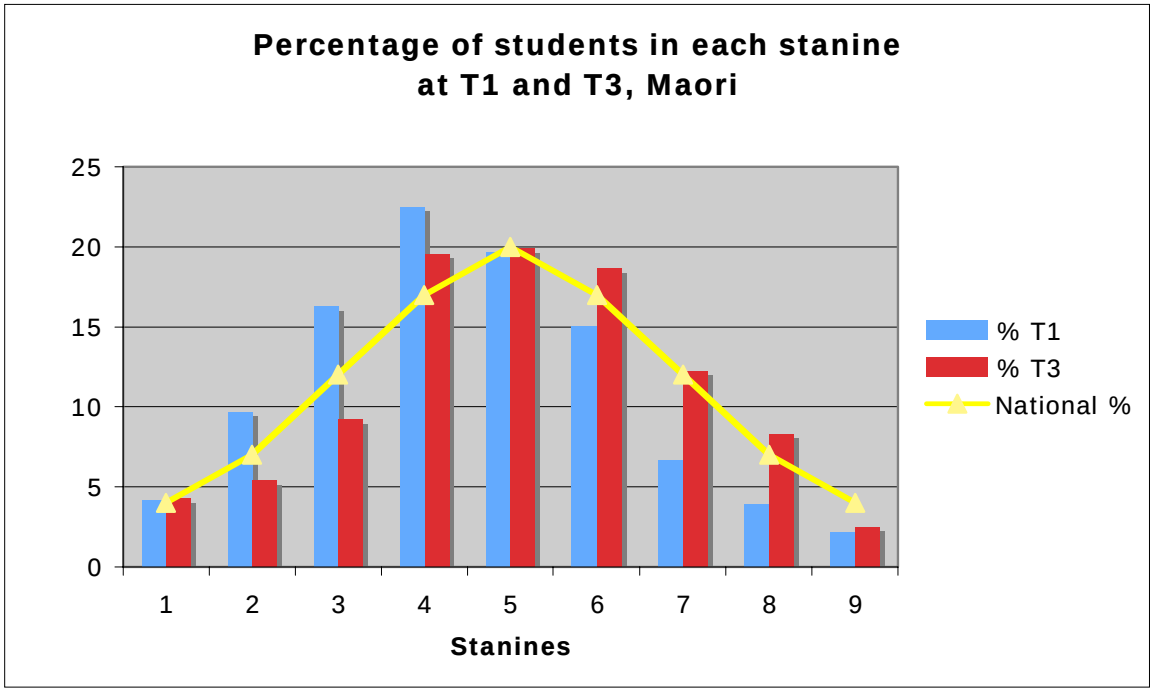

Percentage of students in each stanine at T1 and T3, Maori
25
20
15
10
5
0
1
2
3
4
5
6
7
8
9
Stanines
% T1
% T3
National %

Figure 13: Percentage of students in each stanine at Time 1 and Time 3, Pasifika

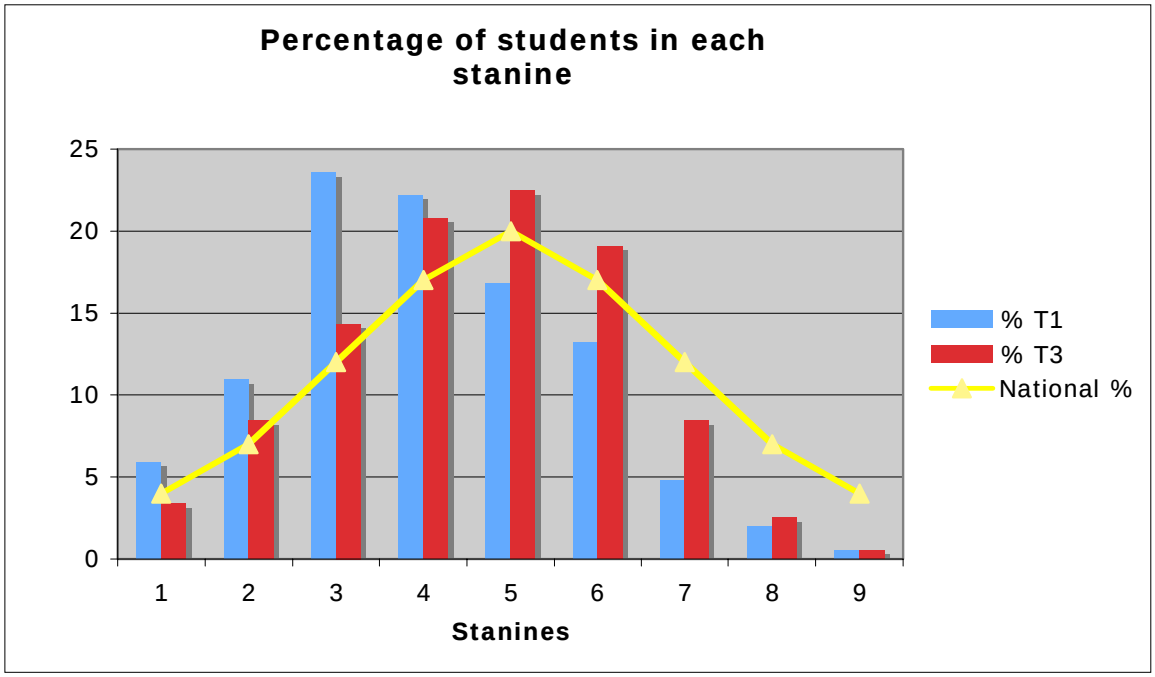
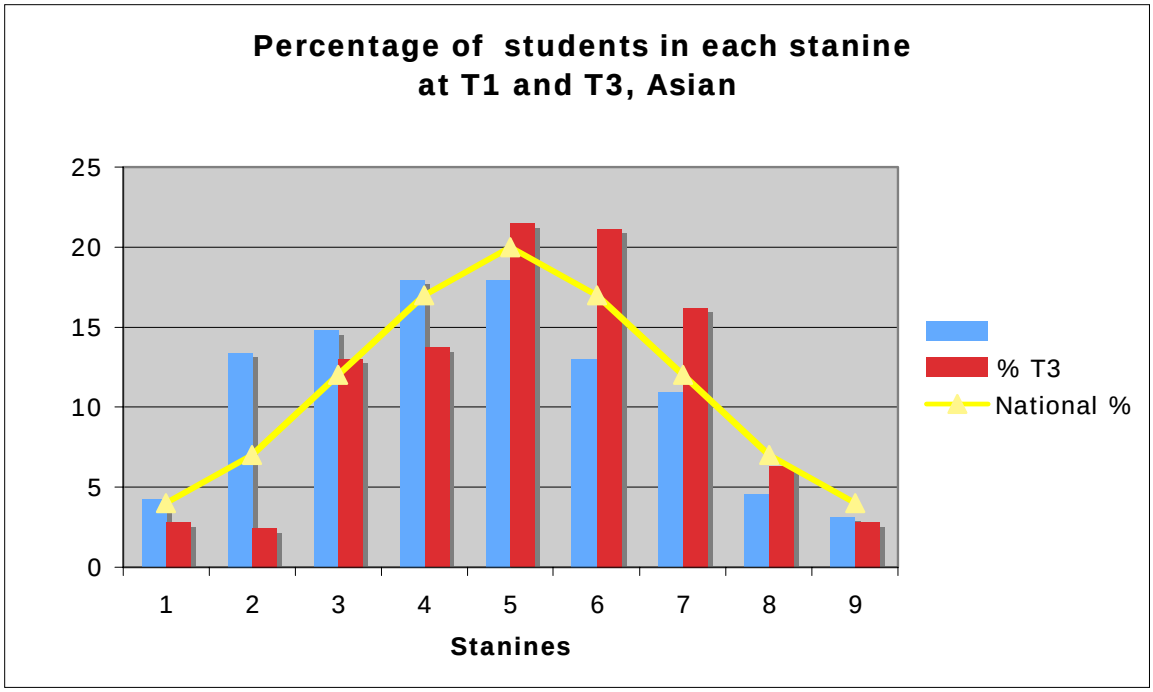


Figure 14: Percentage of students in each stanine at Time 1 and Time 3, Asian



By Time 3, the percentage of Māori students in the lowest three stanines is lower than the national picture (19% compared with 23%) and the percentage of students in the top stanine band is the same as the national picture (23%). The Asian students have a

similar percentage of students in the lower band at Time 3 (18%) but have more than the national picture in the upper band by Time 3 (25% compared with 23%).

The Pasifika cohort had 40.5% in stanines 1–3 at Time 1 and 26% by Time 3. 95% of the students at stanine 1 Time 1are still within the stanine 1–3 band at Time 3. (See Table 24.)

Table 24: Stanine movement from Time 1 to Time 3: Pasifika (percentage of Time 1 stanine)

Pasifika

% of students at T1 stanine		*Stanine T3*								
		1	2	3	4	5	6	7	8	9
Stanine T1	1	**28.57**	28.57	38.10	4.76	0.00	0.00	0.00	0.00	0.00
	2	10.26	**30.77**	25.64	25.64	5.13	2.56	0.00	0.00	0.00
	3	1.19	14.29	**32.14**	35.71	11.90	3.57	0.00	1.19	0.00
	4	0.00	0.00	7.59	**29.11**	45.57	16.46	1.27	0.00	0.00
	5	0.00	0.00	0.00	13.33	**36.67**	38.33	11.67	0.00	0.00
	6	2.13	0.00	0.00	4.26	17.02	**40.43**	31.91	4.26	0.00
	7	0.00	0.00	0.00	0.00	11.76	41.18	**41.18**	5.88	0.00
	8	0.00	0.00	0.00	0.00	0.00	28.57	0.00	**42.86**	28.57
	9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	**0.00**

What shifts do we see when the data is explored at the subtest level?

The raw score subtest data has been analysed year by year and through the lens of ethnicity. For each year group, except year 6, in both the whole cohort and the Māori students' cohort, the greatest mean subtest score shift was in paragraph comprehension. This shift may be due to the fact that the paragraph comprehension subtest has more questions than the other subtests so there is more room for improvement. For year 6, in both the whole cohort and the Māori students' cohort, the greatest mean shift was in vocabulary range. As discussed earlier, this year group had the lowest mean stanine shift from Time 1 to Time 3, so as expected, it had the lowest mean raw score difference. This year group continues to be interesting, as it is the only group where the Māori student cohort had a smaller mean score difference than the year group as a whole.

This subtest analysis, Tables 25 and26 and Figures 15 and 16, could be explored further. For example, why does year 4 have the highest mean score difference (11.33 for the Māori students' cohort and 9.98 for the cohort as a whole) yet the lowest word recognition shift for the cohort as a whole?

Table 25: **Whole cohort STAR subtest score differences**

	Year 3	Year 4	Year 5	Year 6	Year 7
Word recognition	1.27	1.30	0.71	1.23	1.20
Sentence comprehension	0.68	1.82	1.75	0.81	2.04
Paragraph comprehension	3.75	4.99	3.81	0.53	3.59
Vocabulary range	-0.01	1.85	1.86	1.49	1.64
Language of advertising					1.93
Writing style					2.36

Figure 15: Whole cohort STAR subtest score differences, Time 1 to Time 3

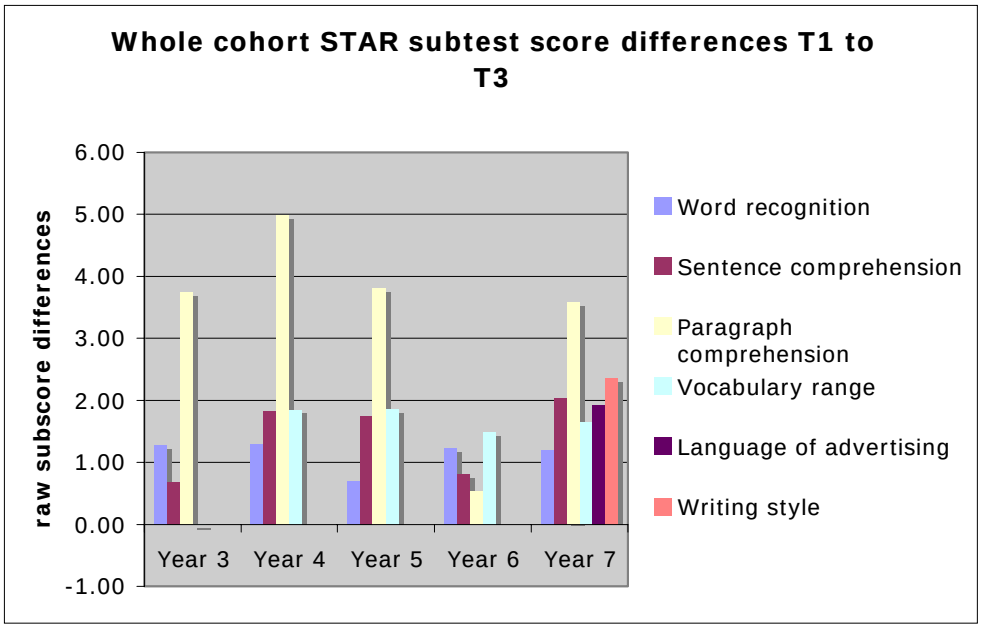
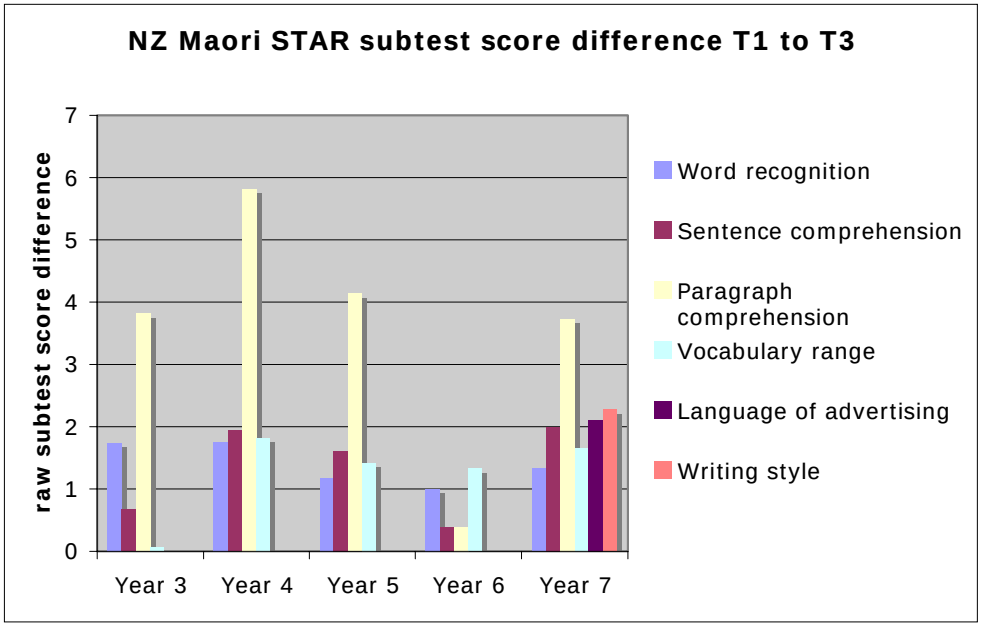


Table 26: Māori student cohort STAR subtest score differences

	Year 3	Year 4	Year 5	Year 6	Year 7
Word recognition	1.74	1.75	1.17	1	1.33
Sentence comprehension	0.67	1.94	1.61	0.39	1.99
Paragraph comprehension	3.81	5.82	4.14	0.39	3.73
Vocabulary range	0.06	1.82	1.42	1.33	1.65
Language of advertising					2.11
Writing style					2.27

Figure 16: Māori student cohort STAR subtest score differences, T1 and T3



Initial findings from the analysis of the STAR data that require further inquiry

It is difficult to know whether the improvements made over the two years by students in stanines 7–9 and/or year 6 are really not as large as the rest of the students, as shown by the STAR results, or whether this low shift is due to the tool's ceiling effect.

Year 6 data often seems to be different from the other year groups in all sorts of different ways. (For example, it shows the lowest overall mean stanine shift but the largest mean stanine shift for the lowest 22.32% and the subtest raw score difference for the Māori student cohort is lower than for the cohort as a whole). Again, without further analysis it is difficult to know whether the reason for this lies with the nature of the teaching or with the tool.

Pasifika students appear to have the lowest mean stanine scores at Time 1 and Time 3 and to have made the smallest shift over the two years. There were 44, 37, 32, 4, and 239 Pasifika students in years 3 to 7, which made it difficult to analyse the data at a subtest and/or year level.

Project response

- The project has decided to use asTTle: Reading for monitoring reading comprehension at a national level for a number of reasons including that there does not appear to be a ceiling effect with this tool (based on the data from the research schools that used asTTle reading in 2005).
- The project directors and the Ministry of Education are exploring material that will inform the facilitation, leadership, and teaching and learning about Pasifika students and literacy teaching and learning. Materials from literacy work, such as

Literacy Leadership, NEMP, and SEMO, are being sourced to build our knowledge so that we can enable schools to be more effective at meeting the diverse needs of Pasifika students.

Writing focus schools

What is the mean score for each year group and how does this compare with the national mean?

The data was analysed to identify the difference in mean total score (aWs) from Time 1 to Time 3 for each year group. The mean shift over the two years was 130.12. The national data (asTTle V4 Manual, 2005, page 3–26)[11] has no two-year shift greater than 50 points. It appears that the national picture shows about one sub-level shift per two years (or less), whereas the students in the project schools made on average a 2.5 sub-level shift in the two years (based on Tables 3.1 {page 3.25} and 4.8 {page 4.17} asTTle V4 Manual, 2005). At Time 3 years 5, 6, and 8 students now have a mean score 1 sub-level above the asTTle national picture and year 7 is 2 curriculum sub-levels above the national picture.

The effect size for each year ranged from 1.27 to 1.41, with a mean of 1.30 (year 3 has not been included in this analysis as these students are not representing all schools' year 3 students).

The descriptive statistics associated with this analysis are shown in Table 27.

Table 27: Mean aWs score at Time 1 and Time 3 for each year group

Year level at Time 1	*N*	*Mean score T1*	*Mean score T3 & curriculum level*	*Difference*	*Effect size 1*[a]	*Year level at T3*	*asTTle national mean for Time 3 year level*[c]	*D b s n*
3	6	336.00	502.50 (3b)	166.50	1.67	4	454	
4	347	374.70	505.84 (3b)	131.14	1.31	5	482	
5	332	433.47	560.15 (3p)	126.68	1.27	6	504	
6	134	459.24	599.94 (3a)	140.70	1.41	7	518	
7	244	476.55	603.18 (3a)	126.63	1.27	8	536	
All	**1063**	**426.87**	**556.99**	**130.12**	**1.30**	**0.27**		

[a] Effect Size 1 is measured using the difference between mean scores and the asTTle score standard deviation of 100
[b] Year 3 statistics are not included in the total as asTTle: Writing was used year 4 – 8 by most schools
[c] National Mean as given by asTTle per year, beginning, end or other point in year unknown

As with the reading data, we wanted to explore whether the "tail" had a greater shift than the cohort as a whole. We identified the shift made by the students in the lowest 20% of each year group. It is interesting that the mean score at Time 1 had a range of 2 curriculum sub-levels (a range of 113 points), whereas the range is a lot smaller at Time 3. The lowest 20% of the year 6 and 7 groups are only one curriculum sub-level below the whole group for their particular year. The lowest 20% of the year 6 students have a

[11] http://www.tki.org.nz/r/asttle/user-manual_e.php chapters 3 and 4

mean score at Time 3 that is equal to the asTTle national mean for that year group. (There is an error margin of 15 points with asTTle).

The mean effect size across the year groups for the lowest 20% of students ranged from 1.87 to 2.17 with an average of 2.05. See Table 28.

Table 28: Mean score at Time 1 and Time 3 for each year for the lowest 20% of distribution at Time 1

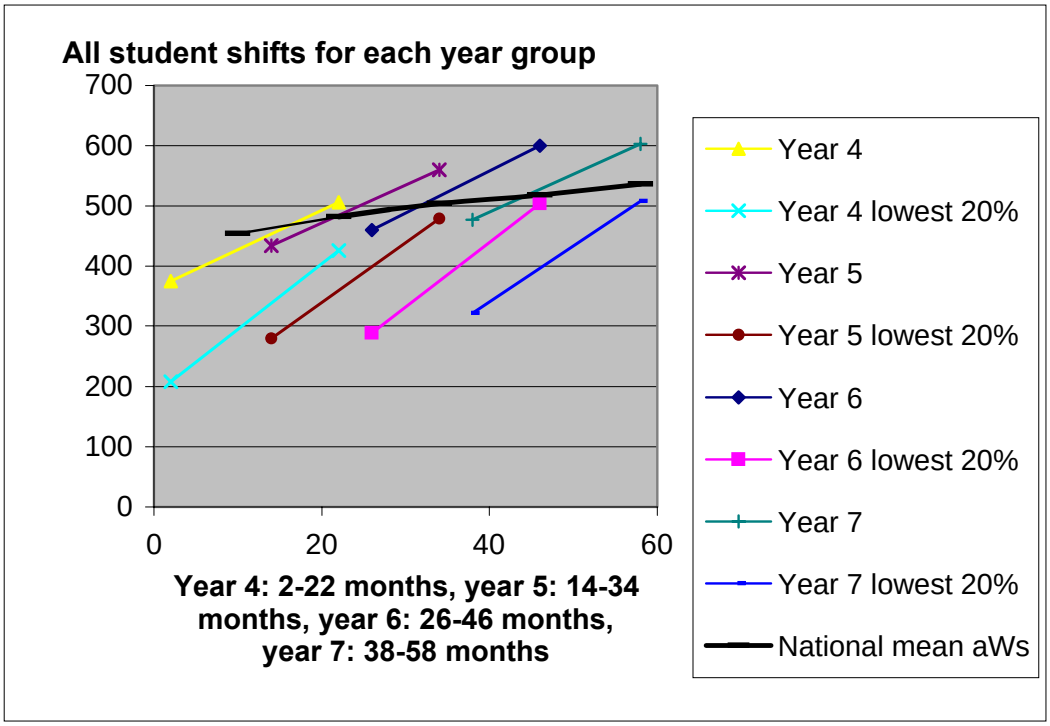
Year Level at Time 1	*N*	*Mean score T1*	*Mean score T3*	*Difference*	*Effect Size 1*[a]	*Year level at T3*	*National Mean for Time 3 year level*[c]	*Difference between mean score T3 and national mean*
3	3	204.00	462.33 (2a)	258.33	2.58	4	454	8.33
4	70	208.57	426.01 (2p)	217.44	2.17	5	482	-55.99
5	69	279.16	478.75 (2a)	199.59	2.00	6	504	-25.25
6	27	288.48	503.30 (3b)	214.81	2.15	7	518	-14.70
7	49	321.24	507.84 (3b)	186.59	1.87	8	536	-28.16
Total 20%	**218**	**266.07**	**471.17**	**205.10**	**2.05**			

[a] Effect Size 1 is measured using the difference between mean scores and the asTTle score standard deviation of 100
[b] Year 3 statistics are not included in the total as asTTle: Writing was used year 4 – 8 by most schools
[c] National Mean as given by asTTle per year, beginning, end or other point in year unknown

It was difficult to know whether to compare the mean score at Time 1 for a higher year group with that at Time 3 for a lower year group. For example, can Time 1, year 5 be compared with Time 3, year 4? Instead, a Twelve-month year was used, with a four-month gap between the Time 3 and the next year's Time 1. Time 1 was put as the second month in the first year and Time 3 was put in as the eleventh month of the second year. This gives us the following x-axis for the graph: year 4: 2–22 months, year 5: 14–34 months, year 6: 26–46 months, year 7: 38–58 months. The national mean was put at Time 3 for each equivalent year group but this is problematic as the asTTle V4 User Manual shows that writing data was collected in November 2000, March 2001, and November 2001 (page 4.9 Table 4.1).

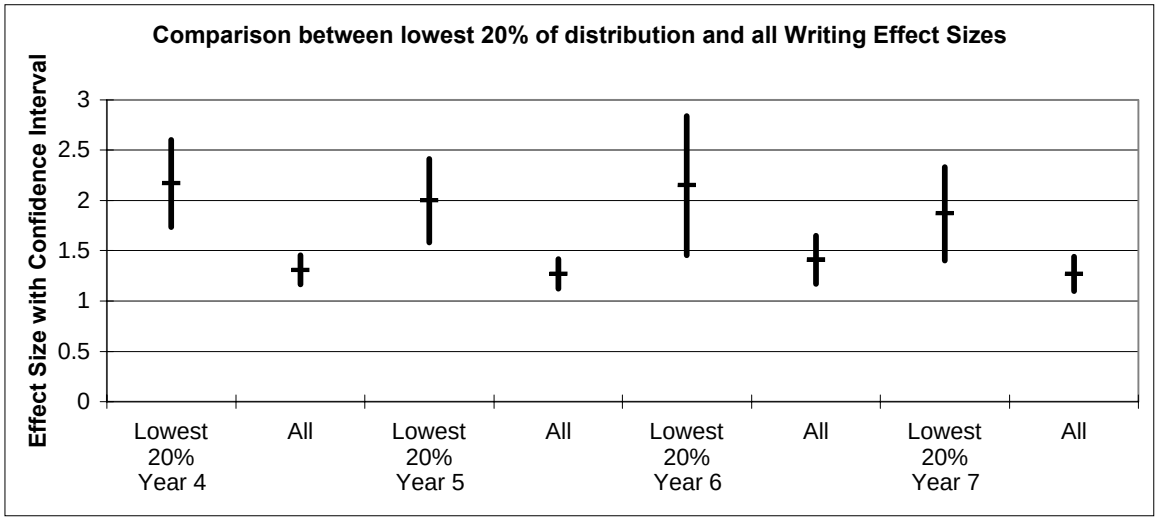
See Figure 17 for shifts from Time 1 to Time 3 for each year group as a whole and for those in the lowest 20% at Time 1.

Figure 17: all students shifts for each year group



It appears that the writing schools within the project have made a significant shift in student achievement for each year group and for the cohort as a whole, but more importantly, they have accelerated the pace of achievement for students in the bottom 20% at Time 1 (as seen by the slope of the lines on the graph). Table 29 shows the confidence limits for the significance of the difference in effect size for the lowest 20% when compared with the appropriate year group as a whole. For years 6 and 7 there is an overlap of possible effect sizes for each group as a whole and the lowest 20% within this group therefore the difference rates of achievement may not be significant.

Table 29: Comparison between the lowest 20% and the whole cohort effect sizes



What shifts do we see when the data is analysed through the lens of gender?

The mean total score (aWs) for all boys puts them one curriculum sub-level than all girls. When the lowest 20% data is analysed the proportion of boys within this subgroup is much larger than the proportion of girls. Boy's pace of achievement is slightly less than girls. See Table 30.

Table 30: Mean score at Time 1 and Time 3 for the lowest 20% of distribution at Time 1

Gender	*N*	*% of whole cohort*	*Mean score T1*	*Mean score T3*	*Difference*
Boys	145	27.6% all boys	261.63	459.28 (2a)	197.66
Girls	73	13.6% of all girls	274.90	494.78 (3b)	219.88
Total 20%	**218**		**266.07**	**471.17**	**205.10**

What shifts do we see when the data is analysed through the lens of ethnicity?

For this analysis, each year group was considered in two ways. Initially, the mean shift and effect size of the shift for each ethnic group was explored. Then the shifts for the lowest 20% of the students at Time 1 for each year group were explored to see if there was a difference in progress by ethnicity. This allowed us to see who was in the lowest 20% and to analyse the shifts they had made.

1. Mean shift and effect size for each ethnic group

For each year group, except year 7, Asian students had the highest mean score at Time 1 and Time 3. Years 5, 6 and 7 cohort of Maori students made the greatest gain in scores over the two years. In year 4 the NZ European / Pakeha cohort had the greatest gain. A number of groups had an effect size larger than 1.30 for the whole cohort.

These groups were: year 4 NZ European / Pakeha students; both year 5 and year 6 Maori students and Pasifika students, and; year 7 Maori students.

See Table 31.

Table 31: Mean score at Time 1 and Time 3 by year

Ethnicity	*N*	*% of whole year group*	*Mean score T1*	*Mean score T3 & curriculum level*	*Difference*	*Effect Size*[a]
Year 4	**340**					**1.31**
NZ European/Pakeha	236	69.41	374.67	514.77 (2p)	140.1	1.4
NZ Maori	62	18.24	360.85	470.84 (2a)	109.99	1.1
Pacific Island	28	8.24	373.54	479.75 (2a)	106.21	1.06
Asian	14	4.12	449.71	559.29 (3p)	109.57	1.1
Year 5	**319**					**1.27**
NZ European/Pakeha	228	71.47	430.51	555.04 (3p)	124.53	1.25
NZ Maori	46	14.42	401.83	544.07 (3p)	142.24	1.42
Pacific Island	24	7.52	438.83	581.13 (3a)	142.29	1.42
Asian	21	6.58	513.14	617.81 (4b)	104.67	1.05
Year 6	**128**					**1.41**
NZ European/Pakeha	95	74.22	456.33	597.25 (3a)	140.93	1.41
NZ Maori	24	18.75	437.79	593.04 (3a)	155.25	1.55
Pacific Island	5	3.91	459.8	569.20 (3a)	109.4	1.09
Asian	4	3.13	567.25	670.25 (4p)	103	1.03
Year 7	**239**					**1.27**
NZ European/Pakeha	162	67.78	491.4	617.19 (4b)	125.79	1.26
NZ Maori	56	23.43	427.96	559.64 (3a)	131.68	1.32
Pacific Island	13	5.44	472	594.77 (3a)	122.77	1.23
Asian	8	3.35	507.63	618.00 (4b)	110.38	1.1
All	**1063**		**426.87**	**556.99**	**130.12**	**1.3**

[a] Effect Size 1 is measured using the difference between mean scores and the asTTle score standard deviation of 100
[b] The **All** row includes all students (not just those in these ethnic groups)

2. Mean shift and effect size for the lowest 20% of students in each year group

Table 32 shows the same sort of analysis for the lowest 20% of each year group. For each year group the effect size of each ethnic group within the lowest 20% students is well above the equivalent effect size for each ethnic group within the whole year group (compare with Table 31). There does appear to be certain ethnic groups over-represented in particular year groups when compared to that year group as a whole.

For example, only year 5 has Asian students in the lowest 20% of students at Time 1 and there are proportionally more Maori students in the lowest 20% of years 6 and 7 than their respective whole year groups. For all year groups, Pasifika students have the highest T3 mean score.

Table 32: Mean score at Time 1 and Time 3 by year for the lowest 20% of distribution at Time 1

Ethnicity	*N*	*% of the lowest 20% for each year*	*Mean score T1*	*Mean score T3& curriculum sub score*	*Difference*	*Effect Size[a]*
Year 4	**67**					**1.31**
NZ European/Pakeha	47	70.15	201.85	432.36 (2p)	230.51	2.31
NZ Maori	14	20.90	214.14	385.43 (2p)	171.29	1.71
Pacific Island	6	8.96	261.83	470.17 (2a)	208.33	2.08
Asian	0					
Year 5	**68**					**1.27**
NZ European/Pakeha	48	70.59	281.85	477.46 (2a)	195.6	1.96
NZ Maori	13	19.12	248.85	471.38 (2a)	222.54	2.23
Pacific Island	6	8.82	316	490.83 (3b)	174.83	1.75
Asian	1	1.47	301	462 (2a)	161	1.61
Year 6	**27**					**1.41**
NZ European/Pakeha	20	74.07	279.25	500.75 (3b)	221.5	2.22
NZ Maori	6	22.22	307.17	506.17 (3b)	199	1.99
Pacific Island	1	3.70	361	537 (3p)	176	1.76
Asian	0	0.00				
Year 7	**48**					**1.27**
NZ European/Pakeha	27	56.25	327	498.48 (3b)	171.48	1.71
NZ Maori	19	39.58	308.89	506.21 (3b)	197.32	1.97
Pacific Island	2	4.17	363	613 (3a)	250	2.5
Asian	0	0.00				
Total 20%	**218**		**266.07**	**471.17**	**205.1**	**2.05**

[a] Effect Size 1 is measured using the difference between mean scores and the asTTle score standard deviation of 100
[b] The **Total 20%** row includes all students in the lowest 20% (not just those in these ethnic groups)

[a] Effect size 1 is measured using the difference between mean scores and the asTTle score standard deviation of 100.
[b] Effect size 2 is the expected measure for one year's development.

Looking at just one sub-group: NZ Māori

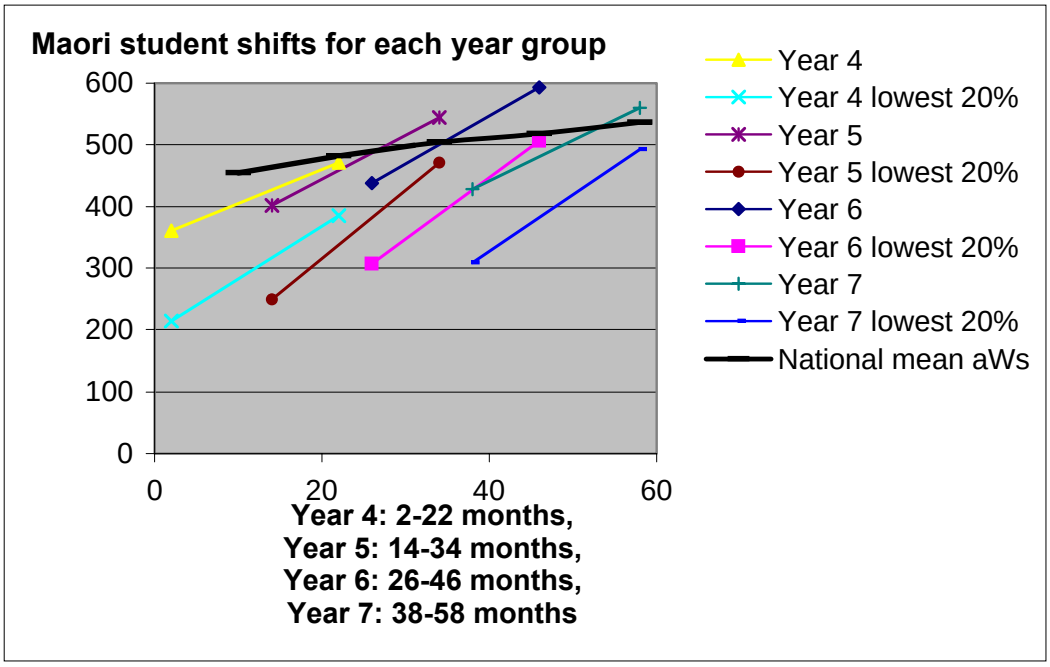
Instead of analysing the data within each year, we chose to look at NZ Māori students across the four year groups. The numbers are very small for the lowest 20% (for example, year 6 has only 3 students) but we need to remember that these numbers are a 10–30 % sample. The mean scores at Time 3 for NZ Māori students in years 5, 6, and 7 (as at Time 1) are above the national mean for their respective year groups. The lowest 20% of the years 5, 6 and 7 NZ Māori students are within 2 curriculum sub-levels of the year 5, 6 and 7 NZ Māori students groups as a whole.

See Table 33 for the descriptive statistics and Figure 18 represents the shifts graphically. Figure 18 has been constructed the same way as Figure 17.

Table 33: Māori student shifts from Time 1 to Time 3 for all years

Year level	*Whole cohort Mean score T1*	*Whole cohort Mean score T3*	*Difference*	*Bottom 20% Mean score T1*	*Bottom 20% Mean score T3*	*Difference*	*National mean score for T3 year level*
4	360.85	470.84	109.99	214.14	385.43	230.51	482
5	401.83	544.07	142.24	248.85	471.38	222.54	504
6	437.79	593.04	155.25	307.17	506.17	199	518
7	427.96	559.64	131.68	308.89	506.21	197.32	536

Figure 18: Writing: Māori students shifts for each year group

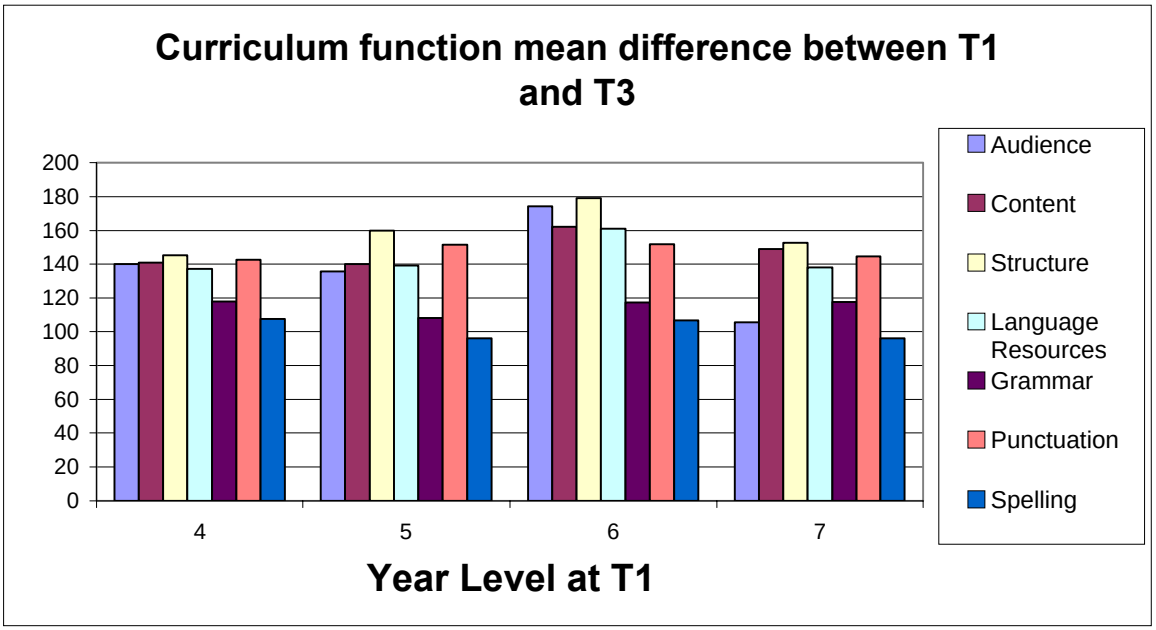


What shifts do we see in the writing content areas for each year group?

Figure 19 shows that the largest shift from Time 1 to Time 3 for each year group was structure. This was the same finding for the case study schools reported in the August 2005 milestone. The focus on structure reflects the anecdotal evidence from facilitators

that this is the shift in teaching focus that they are seeing initially. This evidence also supports one of the research reports (July 2005, chapter 5) that describes the way teachers are able to develop clear learning intentions for students around writing structure so that students know what they are learning and when they have succeeded. It will be worthwhile exploring teacher content knowledge around structure to see whether this links these three findings.

Figure 19: Difference in mean content area score between Time 1 and Time 3



For each year group structure and punctuation had the lowest mean scores at Time 1. At Time 1 the spread among the different content areas was much wider than the spread at Time 3 for all year groups. This can be seen in Figures 20–23.

Figure 20: Year 4 mean content area scores at Time 1 and Time 3

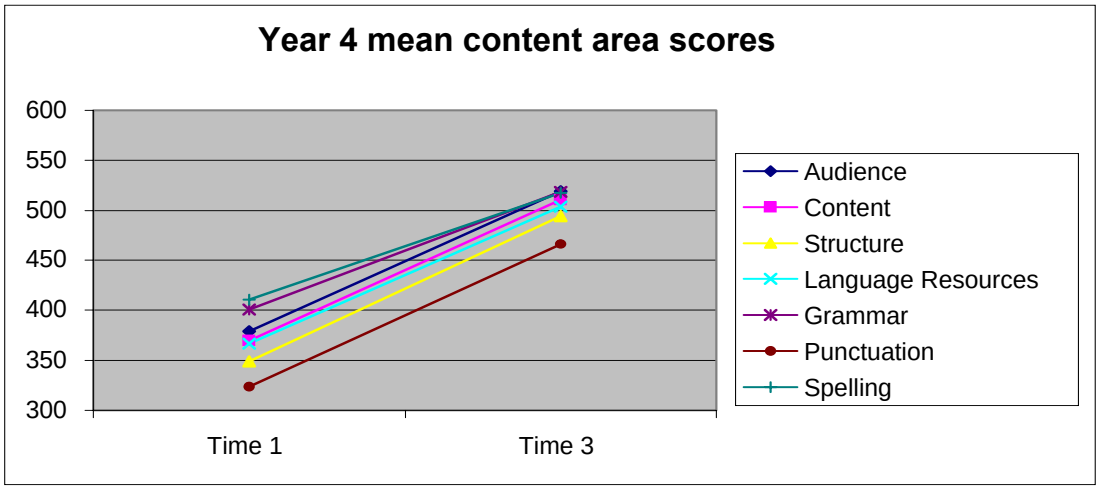


Figure 21: Year 5 mean content area scores at Time 1 and Time 3

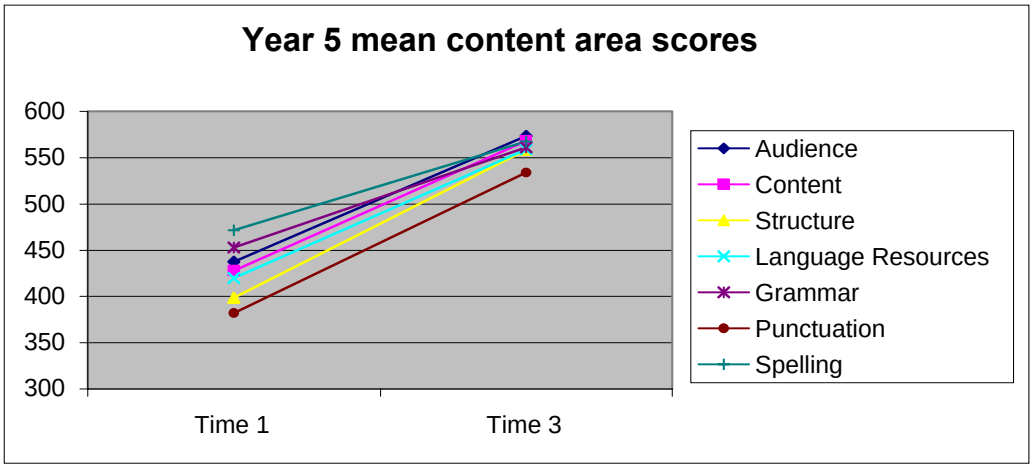


Figure 22: Year 6 mean content area scores at Time 1 and Time 3

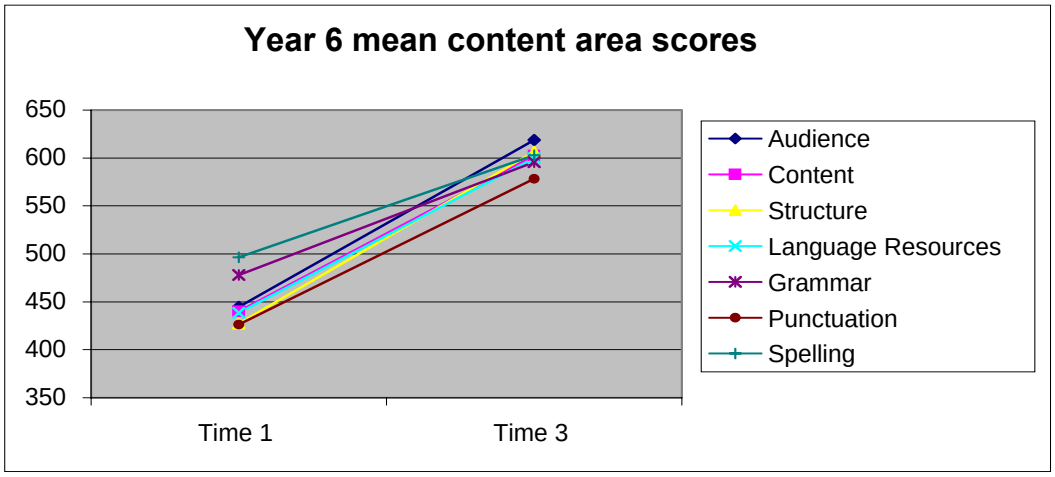
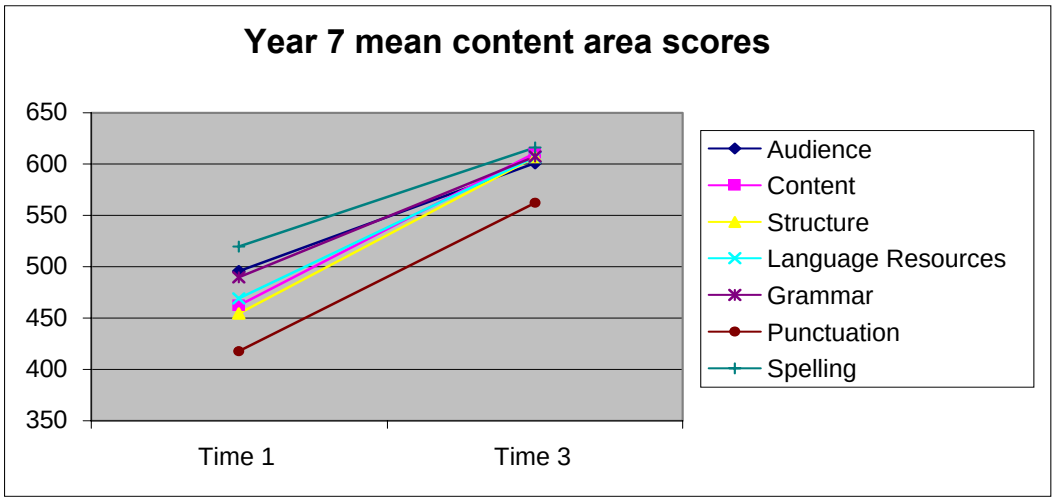


Figure 23: Year 7 mean content scores at Time 1 and Time 3



Initial findings from the analysis of the asTTle data that require further inquiry

Years 6 and 7 mean scores for each year group as a whole and for the lowest 20% of students are not much higher than the years 4 and 5 mean scores. The project needs to explore the teaching at years 6 to 8 to know whether this finding is artefactual or whether it is associated with aspects of teaching.

The Pasifika cohort was not large enough to analyse shifts for each year group. It appears that the Time 3 mean scores for the students in the lowest 20% are higher than those of the other ethnic groups but because the numbers are very small it is not known whether this pattern reflects a true picture.

Project response

- Student achievement data will be collected nationally at Time 1, Time 2, and Time 3 so each year can be explored. For example, are the shifts made in the first year of the project for year 7 students similar or different to the shifts made in the second year for another group of year 7 students (those that were in year 6 at Time 1)?
- asTTle writing scores will be collected from all students in the 2006 cohort. Facilitators will guide schools through the asTTle moderation process as detailed on TKI. If the process is followed, we should be able to trust the data and accept the range of variability in the marking. With greater student numbers and with three data collection points, we will be able to make more sense of what various sub-groups of students are achieving.

Comparing the shifts in achievement for reading and writing

The effect sizes (based on the mean scores) from years 4, 5, and 7 can be compared between reading and writing. (There was not an effect size for year 6 reading, as the STAR test used for year 6 was different to the one used for year 7). By looking at the confidence intervals associated with each effect size (Table 34), it appears that the shift in writing is greater than that in reading for years 5 and 7. The project needs to explore whether this is artefactual or actual.

Table 34: Comparison between the reading and writing effect sizes