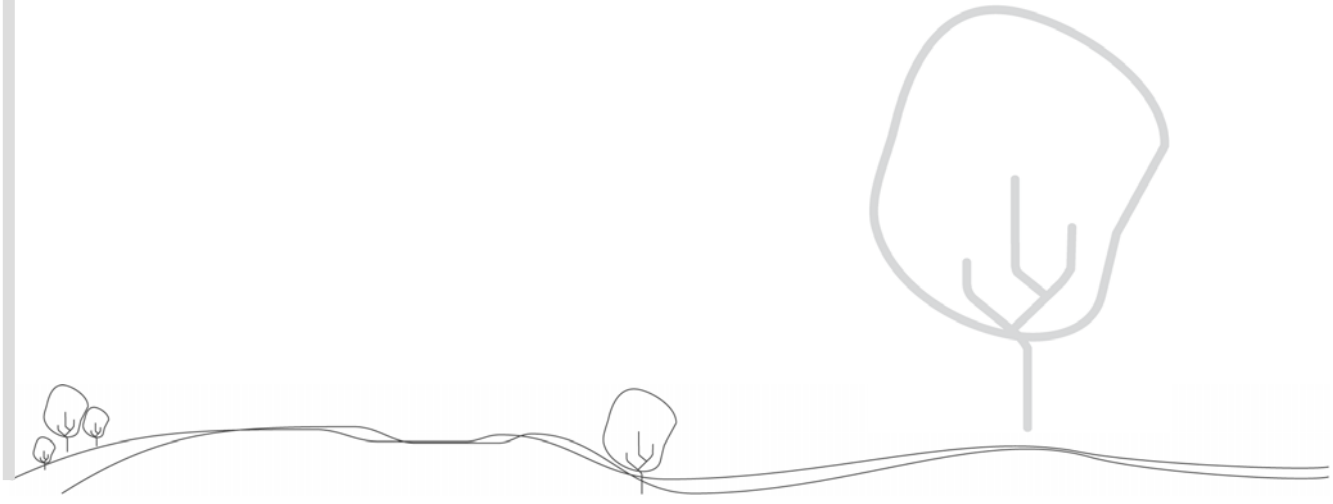
TEACHING & LEARNING
RESEARCH INITIATIVE
NĀU I WHATU TE KĀKAHU, HE TĀNIKO TAKU

PO Box 3237
Wellington, New Zealand
e-mail: tlri@nzcer.org.nz

Enhanced teaching and learning of comprehension in Years 5–8: Otara Schools

Stuart McNaughton, Mei Kuin Lai, Meaola Amituanai-Toloa, and Sasha Farry

2008



Enhanced teaching and learning of comprehension in Years 5–8: Otara Schools

Stuart McNaughton, Mei Lai, Meaola Amituanai-Toloa, and Sasha Farry

Woolf Fisher Research Centre, Faculty of Education, The University of Auckland

2008

Teaching and Learning Research Initiative
P O Box 3237
Wellington
New Zealand
www.tlri.org.nz

Reports from Auckland UniServices Limited should only be used for the purposes for which they were commissioned. If it is proposed to use a report prepared by Auckland UniServices Limited for a different purpose or in a different context from that intended at the time of commissioning the work, then UniServices should be consulted to verify whether the report is being correctly interpreted. In particular it is requested that, where quoted, conclusions given in UniServices reports should be stated in full.

Acknowledgements

This project is the result of a close collaboration between the leaders and teachers in seven schools in Otara and members of the Woolf Fisher Research Centre. We wish to acknowledge the professional expertise of the teachers and leaders in the schools. The achievements described in this report derive from their expert participation as partners. The support and contributions from their school communities, including their boards of trustees and the Otara Boards Forum, are also acknowledged.

Colleagues from the Ministry of Education, both locally and nationally, have been involved at each stage and have been valued members of the collaboration with the schools and the researchers. We wish to thank those colleagues for their high-level policy-based and research-based contributions.

The research and development programme received funding from the Teaching and Learning Research Initiative (co-ordinated by the New Zealand Council for Educational Research), the Woolf Fisher Trust, the seven schools, and the Ministry of Education. This report is to the New Zealand Council for Educational Research, and we wish to thank the director and research team for the opportunity to develop further the research and practice partnerships in South Auckland, and the approach to the management of the research funding which has enabled us to work effectively in an applied setting.

Pauline Te Kare has been critical in administering the work of the centre and the outcomes here owe a considerable amount to her skills. We also acknowledge the work of a number of research assistants, including Jolyn Tay and Angela Kuo, who helped with the data entry and data analysis, often under pressing time constraints, and Maryanne Pale and Azam Riazuddin who supported with document preparation.

The Woolf Fisher Research Centre is a centre of the University of Auckland, supported by Auckland UniServices Limited, and receives funding and support from the Woolf Fisher Trust, The University of Auckland, and Manukau Institute of Technology.

Table of Contents

Tables

Figures

1. Introduction

This study represents a systematic replication of a previous intervention which took place in schools in Mangere from 2003 to 2005. (McNaughton, MacDonald, Maituanai-Toloa, Lai, & Farry, 2006). The contexts and theoretical rationale are the same for the present study as those for the original Mangere study. We have repeated that historical and social context and theoretical framework here. In this first section, however, we briefly summarise the original study and also outline the form and the role of replication in the science represented here.

Replication

Our previous study

In previous quasiexperimental research with a cluster of similar schools in Mangere in South Auckland we have shown that it is possible to teach reading comprehension more effectively and raise achievement levels significantly higher than existing levels, in that study by 0.97 stanine[1] over three years with an overall effect size for gains in stanines of 0.62 (McNaughton et al., 2006). Given that stanine scores are age adjusted, the overall gain of 0.97 stanine indicates that the intervention advanced student achievement across all cohorts by almost a whole school year in addition to the expected national advancements made over three years. In addition, at each year level, achievement was raised significantly higher than the baseline forecast for that year level by up to 1.03 of a stanine.

Moreover, we showed that the intervention required a context-specific analysis of teaching and learning needs (Lai, McNaughton, MacDonald, & Farry, 2004; McNaughton, Lai, MacDonald, & Farry, 2004). That is, specific aspects of comprehension and competing hypotheses about the sources of low levels of comprehension needed to be checked. For example, it was found that accuracy of decoding was not a widespread impediment to comprehension (Lai et al., 2004). Knowing the profile of learning for children led to a planned research and development programme that catered for the specific needs of these particular children.

---

[1] Stanines are normalized standard scores having a mean of five and a standard deviation of about two (Reid & Elley, 1991). They are expressed as a scale of nine units with a low of one and a high of nine. In the PAT manual, stanine nine is described as "superior", stanine seven and eight as "above average", stanine four to six as "average", stanine two and three as "below average" and stanine one as "low". The nine stanine units may be considered as nine categories of reading attainment, making it "highly suitable for interpreting performance on the PAT: Reading" (Reid & Elley, 1991, p. 23).

The evidence from the quasiexperimental design used with the first cluster of schools in Mangere was that an initial phase of cluster-wide and school-based analysis and critical discussion of evidence about teaching and learning was a major component of the intervention, initially contributing up to 0.5 of the shift in the average stanine, although this was in the context of a fixed sequence of phases. That is, when comparing two phases of professional development in sequence starting with critical analysis and then adding professional development focused on developing teachers' instructional practice, the critical analysis phase was associated with marked acceleration in achievement (with continuing although smaller gains when the second phase of intervention targets was added), and that this phase on average produced greater accelerations in achievement than the second phase for the cluster (Lai et al., 2004).

The explanation for the effect of the first phase is that the evidence-based problem-solving process, which involved a professional learning community of teachers, researchers, and policymakers, enabled teachers to fine-tune their teaching practices in specific areas identified by the process. In other words, teachers became inquirers of their own practice, using evidence from student achievement and observations of current instruction to address the identified teaching and learning challenges to raise achievement. Indeed, evidence from classroom observations showed change occurred in the specific areas targeted through the problem-solving process (McNaughton et al., 2006).

This problem-solving approach to professional development has been implicated in other local and international interventions designed to raise student achievement (e.g., Alton-Lee, 2003; Thomas & Tagg, 2005; Timperley, Phillips, & Wiseman, 2003). For example, Timperley et al. (2003) found that schools that engaged more strongly in problem solving around student data had higher student achievement in literacy than schools that did not. This approach has also been proposed as a more effective form of professional development. In their literature review on effective professional development, Hawley and Valli (1999) identified evidence-based problem-solving as an effective form of professional development, perhaps more effective than traditional workshop models. This was confirmed by a recent synthesis which claimed that analysis of data is effective professional development that can be linked to enhanced student learning (Mitchell & Cubey, 2003).

In addition, the success of the intervention appeared to be related to the development and maintenance of professional learning communities, and we found some evidence, especially in a second year, that school-level effects such as leadership and the presence of ongoing participation became important determinants (Lai et al., 2006). This is consistent with research where instructional leadership and the development of professional learning communities as sites for discussing evidence about teaching and learning are implicated in the intervention's success (e.g., Newman, Smith, Allensworth, & Bryk, 2001; Taylor, Pearson, Peterson, & Rodriguez, 2005). Coburn (2003) even suggests that sustainability of an intervention requires support from multiple levels of the system, including communities of teachers and school leaders discussing practice within and across schools. This is consistent with the results of the first (Mangere) intervention. The research–practice collaboration is part of the wider strategy for sustainability in local

initiatives, where schools form strategic partnerships, in this case with research institutions, to create sustainable area-wide interventions to raise student achievement.

In the second year a professional development programme, focused on specific aspects of the teaching of reading comprehension, was added to continued problem solving by the communities. Unlike some other interventions, the specific practices were highly contextualised, developed directly from the profiles of teaching and learning identified in the first phase. Observations conducted in classrooms in the second phase showed that targeted aspects of instruction, such as increasing the focus on vocabulary teaching, instruction aimed at increasing paragraph-level comprehension, and developing students' awareness of tasks and strategies were taking place. The overall gain in the second phase was 0.35 stanine.

The explanation for gains in the second phase was that professional development aimed at identifying and fine-tuning specific practices was needed in addition to the processes implemented in the first phase. Despite the substantial gains in the first phase (0.47 stanine), they were not sufficient to achieve the goal which the school communities had set of parity with national distributions. Moreover, there were cohorts which made the same or even higher gains in the second phase and were only then approaching national levels in their classrooms. Despite greater variability in the second phase, and lower overall gains, it did not appear that the professional development focused on specific instructional practices was of lesser significance per se. One interpretation of the results was that gains following or in addition to analysis are harder to achieve.

Phase Three continued the critical discussion of Phase One and the teaching targeted in Phase Two. Further professional development did not occur, but further components designed to build the critical discussion around evidence through the professional learning communities within and across schools were added. The indicators for these attributes in the third phase included the continued involvement of schools in the process of critical discussion and the designing, implementing, and collectively reporting of classroom-based projects in a teacher-led conference. In general, there was a high rate of engagement by teachers as well as leaders in the conference. The topics for projects were theoretically based, the teachers gathered and reported on evidence, they adopted an analytic stance to that evidence and they related their analyses to the patterns of student learning and teaching in their classrooms. The evidence from the achievement data was that the intervention was sustained in the third year. The rate of gain increased in the third phase (0.51 stanine) compared with the second.

Good science requires replications

In quasiexperimental research the need to systematically replicate effects and processes is heightened because of the reduced experimental control within the design. This need is specifically identified in discussion about alternatives to experimental randomised designs (Borko, 2004; Chatterji, 2005; Raudenbusch, 2005). For example, McCall and Green (2004) argue

that in applied developmental contexts, evaluation of programme effects requires a variety of designs including quasiexperimental, but our knowledge is dependent on systematic replication across site analyses. Replication across sites can add to our evaluation of programme effects, particularly when it is inappropriate or premature to conduct experimental randomised designs. Such systematic replication is also needed to determine issues of sustainability (Coburn, 2003). Coburn argues that the distribution and adoption of an innovation are only significant if its use can be sustained in original and subsequent schools.

There is debate within New Zealand about how influential school-based interventions focused on teaching practices can be in raising achievement (Nash & Prochnow, 2004; Tunmer, Chapman, & Prochnow, 2004). Given the significance of these counter-arguments to policy directions, it is important to add to the evidence of teacher effectiveness (Alton-Lee, 2003). Therefore, whilst results from the collaboration with the Mangere cluster of schools suggested the significance of the three-phase intervention, with statistically significant improvements in student achievement across all year levels and schools (McNaughton et al., 2004), it is important that these results be replicated to provide further evidence of the impact of the intervention.

In the research with the Mangere cluster of schools, there were inbuilt replications across age levels and across the cluster of schools within the quasiexperimental design format (McNaughton et al., 2004). However, there are possible competing explanations for the conclusions of the cluster-wide results which are difficult to counter with the quasiexperimental design. One is that the immediate historical, cultural, and social context for these schools and this particular cluster meant that an unknown combination of factors unique to this cluster and these schools determined the outcomes. For example, it might be that the nature of students changed in ways that were not captured by the general descriptions of families and students. Or, given that the immediate history included a number of initiatives such as ECPL (Early Childhood Primary Links) and AUSAD (Analysis and Use of Student Achievement Data) (Annan, 1999), the schools were developing more effective ways of teaching anyway.

In this report we describe a systematic replication in a second cluster of schools, in Otara, in South Auckland. The Otara cluster of schools shared significant similarities with the Mangere cluster in terms of geographical location and school and student characteristics. A similar number of schools were involved, they were from neighbouring suburbs, and they had similar proportions of ethnic groups and a similar history of interventions, in that both clusters had been involved in similar government schooling-improvement initiatives in the years leading up to the research. The research–practice collaboration, as was the case in the original research intervention, involved a three-phase research and development sequence designed to improve the teaching of reading comprehension in the middle school years to raise student achievement. The collaboration involved researchers working with teachers and school leaders of urban schools with the lowest employment and income levels in New Zealand, serving largely Māori and Pacific communities. The replicated process was collaborative, staged, analytic, theoretically intense, and culturally located (see McNaughton et al., 2004, for details).

Yesterday was too late?

In 1981, Peter Ramsay (Ramsey, Sneddon, Grenfell, & Ford, 1981) and his colleagues at the University of Waikato completed a study of the schools in South Auckland. The title of their report was *Tomorrow may be too late.* They argued that there was an impending crisis created by "educational disadvantage suffered by most school-aged students in Mangere and Otara" who were "achieving well below their chronological age" (p. 41). They concluded with "a plea for urgency as the needs of the children of Mangere and Otara are very pressing. Tomorrow may be too late!" (p. v).

The gap in achievement between Māori and non-Māori children in mainstream schools is not a recent phenomenon. Earlier reports, such as the Currie (1962) and Hunn (1961) reports on education in the 1950s, had identified this difference as important and as urgently in need of a solution (see also Openshaw, Lee, & Lee, 1993). The long-standing issue on the "problem" for Māori students is important to note, because some commentaries suggest it is relatively recent and can be linked to changes in methods of teaching reading and writing which began in the 1960s (Awatere-Huata, 2002; Nicholson, 2000).

Yet the historical picture is not entirely bleak. There is evidence that in the colonial period, there were times when Māori children outperformed other children in some schools. Some evidence for this can be found in the Survey of Native Schools for 1930 (Education Gazette, 1930; see also McNaughton, 2000).

The sense of crisis that Ramsay expressed for the sake of children, communities, and families is also present in reports from other countries (Snow, Burns, & Griffen, 1998). The need is identified for communities who have, relative to the mainstream communities, less economic and political power, whose children are considered to be "minorities". But there has been little evidence that the crisis is able to be solved in schools. In the United States, Borman (2005) shows that national reforms to boost the achievement of children in low-performing schools serving the poorest communities have produced small gains in the short term (of the order of effect sizes of less than 0.20), but that after seven years, in those few schools that sustain reforms over a long period, the effects increase (estimated to be around effect sizes of 0.50). When considered across the country, while some achievement gains have occurred, they have typically been low and need to be accumulated over long periods of time.

At a more specific level, some studies from the United States have shown that clusters of schools serving "minority" children have been able to make a substantial difference to the achievement of children. In one set of studies (Taylor, Pearson, Peterson, & Rodriguez, 2005), researchers who intervened in high-poverty schools with carefully designed professional development research and development found small cumulative gains across two years too. This study and others pointed to important school-level factors that must be in place in order for all children to achieve at high levels in reading. Summarising these, Taylor et al. (2005) noted six key elements: improved student learning, strong leadership building, strong staff collaboration, ongoing professional development, sharing student assessment data, and reaching out to parents. In these studies there

is evidence that achievement can be effected, and in the case of studies such as Taylor et al. (2005), that small gains over two years could be attributed to these characteristics.

The days after Ramsay's tomorrow

Where does such offshore evidence leave the schools of South Auckland, which, according to Ramsay, had already received substantial additional resources by the early 1980s? There is little evidence that Ramsay's concern led to immediate changes. The evidence from both national and international comparisons suggests that by the beginning of the 1990s, the children in decile 1 schools, and more generally children who were Māori and Pasifika, were still not achieving as well as non-Māori and non-Pasifika children in reading comprehension. The reading comprehension comparisons across 32 countries in the International Association for Evaluation of Educational Achievement study (IEA, 1992) provided stark evidence of what came to be called a "long tail" in the distribution of achievement. The problem was that while in general New Zealand continued to have high average achievement, and the best students in New Zealand were superior to other students in the world, Māori and Pasifika children were over-represented in the "long tail" (Elley, 1992; Wagemaker, 1992).

In New Zealand the recognition of the distribution problem, as well as other research developments, has had an effect. Reports by a Literacy Task Force (1999) and a Literacy Experts Group (1999) contributed to a national policy shift, which was implemented in the National Literacy and Numeracy strategy. The policy shift promoted concerted professional development and research practice development which was focused on Years 1–4 and Māori and Pasifika children, especially those in decile 1 schools.

Associated with this policy and practice shift, there is now evidence from the national educational monitoring project (NEMP) and renorming exercises that changes in the targeted areas have occurred (Elley, 2005). The news is positive for the early stages of literacy instruction. From NEMP, the one area in literacy achievement where there are clear changes is in reading decoding, both accuracy and fluency (Flockton & Crooks, 2001). Their second cycle of assessments of reading showed that the percentages of children reading below age level in reading accuracy at Year 4 had reduced markedly from 1996 to 2000, from around 20 percent to around 13 percent. Little improvement occurred for Year 8 children in oral reading (Flockton & Crooks, 2001). A recent renorming of standardised assessments at Year 1 (6 years) conducted in 2000 also suggests that knowledge of letters and sounds has improved (Clay, 2002).

These increases in oral reading accuracy were found to have been maintained in the third (2004) cycle of assessments at Year 4. Further notable increases in accuracy were found for the Year 8 children, with only around 11 percent at both year levels now reading below age level (Crooks & Flockton, 2005). The breakdown of gains in 2000 and 2004 suggest that reading accuracy had improved at similar rates at Year 4 for both Māori and Pākehā children (Flockton, 2003). But by

2004, the analyses showed substantial reduction at Year 4 in the gap between Pākehā and Māori students (see further comment at nemp.otago.ac.nz/forum_comment/ 2004).

Research-based interventions using experimental designs have shown that the gaps at this early stage can be reduced considerably. We also know many of the characteristics of effective teaching at that early stage. For example, in the *Picking up the Pace* research with Māori and Pasifika children in decile 1 schools in Mangere and Otara, their typical achievement was two stanines below average levels in areas of decoding after a year at school (Phillips, McNaughton, & MacDonald, 2004). A research-based intervention used professional development with teachers and teacher leaders to increase effectiveness in areas of reading and writing, including specific phonics instruction. Where teaching approaches were fine-tuned to solve children's confusions and to make the purpose of classroom activities more obvious, and higher expectations about achievement were developed through evidence-based analyses of progress, the children's achievement was raised to close to the national distribution (see Phillips et al., 2004). In some areas, such as alphabet knowledge, their progress was as good as or better than typical progress; in others, for example, progress through text levels, they closely approximated typical progress; but in one area, generalised word recognition, they were still noticeably below average levels.

"Tomorrow" is still the same for reading comprehension

These indicators of progress are cause for some celebration, given the urgency signalled in Ramsay's report, and the seemingly intractable nature of the teaching difficulty over decades. But the news has not all been good. For reading comprehension, little appeared to have changed for Māori and Pasifika children in low-decile schools over the period in which the decoding changes occurred, as we will show below. The NEMP data indicate increases in levels of comprehension in Year 4 from 1996–2000, but the breakdown of the achievement patterns suggests a substantially wider disparity between Māori and non-Māori in comprehension both at Year 4 and at Year 8. Furthermore, for children in low-decile schools, gaps in comprehension increased both at Year 4 and at Year 8 (Flockton, 2003).

In 2004 the gains in oral-reading accuracy which were maintained in the third cycle of assessments were not, however, matched by similar gains in reading comprehension for the total group of students at either Year 4 or Year 8. The detailed comparisons suggest that the gaps in oral reading accuracy between Māori and Pasifika students and Pākehā students which had closed between 1996 and 2000 reduced further in 2006. But this was not matched in comprehension (Crooks & Flockton, 2005). Commentaries on this 2004 report note that Māori children performed well in decoding, but there were large differences in favour of Pākehā in aspects of comprehension (NEMP, 2004). These differences were apparent for Pasifika children too, and they were apparent for decile 1–3 schools when compared with other decile groups (Crooks & Flockton, 2005).

This is true also of at least some of the schools in South Auckland. When we completed a baseline profile of a cluster of schools in Mangere (as described earlier in the introduction), we found that across schools and year levels, achievement in reading comprehension was relatively flat at around stanine 3, and something like two years below what would be expected as average progress nationally (Lai et al., 2004). This was also found in a baseline profile for the present similar cluster of schools in Otara as described further in this report. They too were, on average, around stanine 3 across year levels and across schools (see Lai et al., 2006 for full details).

What we now know is that even if we achieve a dramatic change in teaching early reading, it does not necessarily mean that the gap reduces further on up the system. Experimental demonstrations specifically targeting the teaching of phonics also tend to show very limited transfer to comprehension (Paris, 2005). Recent national data from the Assessment Tools for Teaching and Learning (AsTTle) project across multiple dimensions of reading comprehension confirm the NEMP picture of large differences between Māori and Pasifika children and other children which are stable across decile levels, despite significant trends of higher achievement from lower to higher decile-level schools (Hattie, 2002).

These comparisons need to be treated with an important qualification. The broad description of these disparities can mask important aspects of the literacy development and achievement of children in so-called "minority" groups. The conventional indicators of school literacy represent some of what children learn about literacy. But children who are in communities which have low employment, low incomes, and minority cultural and language status have engaged in a range of literacy and language activities, some of which might be quite different from mainstream children. Their knowledge, therefore, may not be well represented in tests of conventional literacy practices, especially at the beginning of schooling (McNaughton, 1999; Snow et al., 1998) and as they move into the middle school levels.

Here, it is important to note that there is an urgent challenge which has strategic importance to all of New Zealand. Students now need greater ranges and levels of knowledge and skills for postsecondary school study and for employment than ever before. Education is increasingly important to the success of both individuals and nations (Darling-Hammond & Bransford, 2005). Overrepresentation of particular groups in low achievement bands is not acceptable at individual, community, or national levels, no matter what the proportion of the population. It is a pressing matter of cultural, political, constitutional (Treaty of Waitangi), ethical, economic, and educational significance that we develop more effective forms of instruction for these students. It is worth noting that by 2021, Māori children will comprise 28 percent and Pasifika children 11 percent of all the under-15-year-olds in New Zealand (Statistics New Zealand, 2002). In Mangere and Otara schools, children from these communities already make up over 90 percent of many school rolls.

There is an additional dimension to that challenge. Many of the children in the South Auckland schools have a language other than English as their home language. Yet language development for these children is not well understood. In the context of bilingual instruction, for example, and the

relationships between development in two languages and two systems of literacy, little is known about biliteracy development and relationships with literacy and literacy instruction (Sweet & Snow, 2003).

Twenty-five years after the Ramsay report, we can report in this study important gains in reading comprehension in Year 4–8 students in decile 1 schools in Otara. This report describes the science of these changes and documents the research and development programme that had taken place. However, the science is closely bound up with a policy context of associated changes in practices. It is likely that without the policy context, the science involved in developing more effective instruction would have achieved less. The results reported here need to be considered with this policy context in mind (Annan & Robinson, 2005).

In addition, the research is located in a particular historical context of school-based interventions. One is the landmark study *Picking up the Pace* (Phillips, McNaughton, & MacDonald, 2001). As noted above, this focused on instruction in the first year, and set out to examine the separate and combined effects on children's achievement of providing co-ordinated professional development to teachers in early childhood settings, and to teachers of children in their first year of schooling. Since the success of that project, which was completed in 2001, further professional development for Year 1 teachers has occurred, based on the practices identified in the research, and the programme in some schools has been extended through to Year 3.

That study and its further development were part of a much broader project initiative, Strengthening Education in Mangere and Otara (SEMO), which aimed to raise the achievement levels of children in these two areas. SEMO's general aim was to strengthen schools in the area and to enhance children's learning opportunities, particularly in literacy, by enhancing the work of early childhood and primary teachers who were providing literacy programmes. SEMO was succeeded by a further policy and practice development in Mangere and Otara, Analysis and Use of Student Achievement Data (AUSAD). This project is located within that government-funded school improvement initiative. The goal of AUSAD is to offer high-quality learning environments to raise achievement. This is done by using student achievement information to inquire into the nature of the underachievement, to test competing explanations of its cause, and to monitor the impact of teachers' decisions about how to intervene. In short, the focus is on developing the inquiry skills of teachers to improve school practices and student learning outcomes. The initiative comprises a number of interventions focusing on improving literacy and numeracy achievement (e.g., the Third Chance programme aimed at improving literacy in Years 1–3).

Reading comprehension

Recent commentaries identify a major theoretical challenge facing literacy instruction. Now that some of the pressing issues in beginning reading instruction (but by no means all) have been resolved, the challenge concerns the teaching of reading comprehension. Higher levels of reading comprehension and related areas of critical thinking are central to the purposes of contemporary

schooling, and are part of the education priorities and key competencies that have been set for New Zealand education (Ministry of Education, 2005). But there is a critical need for research into instruction that enhances comprehension, and into interventions that enable schools to teach comprehension effectively. The most recent reviews of relationships between research and practice note that overall evidence of teacher effectiveness is limited, and that research has not impacted greatly on effective comprehension instruction (see Block & Pressley, 2002). Similarly, the RAND (Research and Development) reading study group, which was set up in 1999 by the US Department of Education's Office of Educational Research and Improvement to identify the most pressing needs for research in teaching reading, has concluded:

> We have made enormous progress over the last 25 years in understanding how to teach aspects of reading. We know about the role of phonological awareness in cracking the alphabetic code, the value of explicit instruction in sound–letter relationships, and the importance of reading practice in producing fluency.… The fruits of that progress will be lost unless we also attend to issues of comprehension. Comprehension is, after all, the point of reading. (Sweet & Snow, 2003, p. xii)

The challenges to teaching effectively have been identified (Pressley, 2002; Sweet & Snow, 2003). One is the need to build on the gains made in research about instructional practices for beginning literacy. A second is to do with knowledge transfer, a failure to turn all that we know about comprehension and comprehension instruction into generally more effective teaching. These needs are particularly significant for schools serving culturally and linguistically diverse populations in low-income areas (Garcia, 2003).

As noted above, on average, students in the middle years of school in New Zealand have high levels of reading comprehension, judged by international comparisons; however, there are large disparities within the distribution of achievement. These are between children from both Māori and Pasifika communities in urban schools with the lowest employment and income levels, and other children (Alton-Lee, 2004). These findings highlight the need for instructional approaches that enable teachers to develop, use, and sustain effective teaching of reading comprehension with culturally and linguistically diverse students. For Pressley (2002), this challenge represents an application problem.

We know a lot about what students need to be able to do, which includes such things as regulating strategy use, and we know a lot about specific instructional effects, such as the need for explicit strategy instruction. What he claims we have failed to do is translate that knowledge into widespread usage with known effects. While Sweet and Snow (2003) echo this claim in their RAND summary of reading comprehension instruction, they also argue that there is yet more to be known about specific teaching and learning relationships, especially in the context of diverse readers, diverse text types, and diverse instructional contexts.

Generally, there is considerable consensus around what students need to learn, and what effective teaching looks like. In order to comprehend written text, a reader needs to be able to decode accurately and fluently, and to have a wide and appropriate vocabulary, as well as appropriate and

expanding topic and world knowledge, active comprehension strategies, and active monitoring and fix-up strategies (Block & Pressley, 2002; Pressley, 2002). So it follows that children who are making relatively low progress may have difficulties in one or more of these areas. The consensus around teaching effectively identifies attributes of both content (curriculum) and process (Taylor et al., 2005). For the middle grades, these include instructional processes in which goals are made clear, and which involve both coaching and inquiry styles that engage students in higher level thinking skills. Effective instruction also provides direct and explicit instruction for skills and strategies for comprehension. Effective teaching actively engages students in a great deal of actual reading and writing, and instructs in ways which enable expertise to be generalisable and through which students come to be able to self-regulate independently.

In addition, researchers have also identified the teacher's role in building students' sense of self-efficacy and, more generally, motivation (Guthrie & Wigfield, 2000). Quantitative and qualitative aspects of teaching convey expectations about students' ability which affect their levels of engagement and sense of being in control. These include such things as text selection. Culturally and linguistically diverse students seem to be especially likely to encounter teaching which conveys low expectations (Dyson, 1999). There are a number of studies in schooling improvement which have shown how these can be changed. In general, changes to beliefs about students and more evidence-based decisions about instruction are both implicated, often in the context of school-wide or even cluster-wide initiatives (Bishop, 2004; Phillips et al., 2004; Taylor et al., 2005).

Just as with the components of reading comprehension, it follows that low progress could be associated with teaching needs in one or more of these areas. Out of this array of teaching and learning needs, those for students and teachers in any particular instructional context will have a context-specific profile. While our research-based knowledge shows that there are well-established relationships, the patterns of these relationships in specific contexts may vary. A simple example might be whether the groups of students who make relatively low progress in a particular context, such as a cluster of similar schools serving similar communities, have difficulties associated with decoding, or with use of strategies, or both, and how the teaching that occurs in those schools is related to those difficulties.

Several hypotheses are possible for the low levels of reading comprehension which are tested in the following research. One is that children's comprehension levels are low because of low levels of accurate and fluent decoding (Tan & Nicholson, 1997). A second is that children may have learned a limited set of strategies; for example, they may be able to recall well, but are weaker in more complex strategies for drawing inferences, synthesising and evaluation; or they may not have been taught well enough to control and regulate the use of strategies (Pressley, 2002). Other possible contributing reasons might be more to do with language: that is, children's vocabulary may be insufficient for the texts used in classroom tasks (Biemiller, 1999); or they may be less familiar with text genres. Well-known patterns of "Matthew effects" may be present in classrooms, where culturally and linguistically diverse children receive more fragmented instruction focused on decoding or relatively simple forms of comprehending, or receive

relatively less dense instruction, all of which compounds low progress (McNaughton, 2002; Stanovich, West, Cunningham, Cipielewski, & Siddiqui, 1996). There is also a set of possible hypotheses around whether the texts, instructional activities, and the pedagogy of the classroom enable cultural and linguistic expertise to be incorporated into and built on in classrooms (Lee, 2000; McNaughton, 2002). But each of these needs to be checked against the patterns of instruction in the classrooms in order for the relationships to be tested.

This approach focuses on the need to understand specific profiles. Rather than test in an ad hoc way the significance of certain teaching and learning relationships, what we did in the study was to test a package of targeted relationships. These are relationships initially identified through a process of profiling both learning needs and patterns of existing instruction. The analysis is aimed at adding further to our research-based knowledge of relationships between teaching and learning in specific contexts, and thereby contributing to the research and application challenges signalled by Pressley (2002) and Sweet and Snow (2003).

We assume in this profiling that while much is known, there are still some areas where we need more knowledge and analysis. This need is pressing in the context of cultural and linguistic diversity. An example in our contexts is the role of activation and deployment of background knowledge. A theoretical argument is often made that instruction needs to incorporate more of the cultural and linguistic resources that minority children bring to classrooms (McNaughton, 2002). But complementing this is another argument: that students need to develop more awareness of the requirements of classroom activities, including the relationships between current resources and classroom activities (McNaughton, 2002). While the general hypothesis of the significance of background knowledge is well demonstrated in controlled studies of reading comprehension (Pressley, 2002), the particular idea of teachers being able to incorporate this, and balancing it with enhancing awareness of classroom requirements, has not been well tested.

In the following study we draw on known properties of effective comprehension and on known relationships between types of instruction and learning outcomes. But we apply this knowledge in an intervention context. Within that context, we test the significance of the assumed relationships and features of teaching and learning. Because the context includes substantial numbers of children for whom English is a second language and who come from diverse cultural backgrounds, this is also a context for discovering new learning needs and new relationships between teaching and learning.

Professional learning communities and critical analysis of evidence

A previous study, focused on literacy achievement over the transition to school, demonstrated substantial gains across a cluster of 12 decile 1 urban schools with primarily Māori and Pasifika students (Phillips et al., 2001). Among other things, the programme involved intensive collection and analysis of achievement data within schools and across a group of schools. Instructional

approaches were modified to impact more strongly on increasing student engagement and teaching effectiveness around agreed goals. Team leaders within schools led professional communities. While the initial development took place within schools over six months, the programme has now been in place in schools for several years. Follow-up research has indicated that those schools which maintained and built on these processes, through a professional learning community focused on teaching and learning, have increased student achievement over time (Timperley et al., 2003).

The features of these learning communities appear similar to those described by Newman et al. (2001). They identify high "instructional programme coherence", a necessary condition for improvements in student achievement that are more likely to be sustained over time. These authors define high instruction coherence as "a set of interrelated programmes for students and staff that are guided by a common framework for curriculum, instruction, assessment, and learning climate and that are pursued over a sustained period" (p. 229). The elements suggested which are crucial to high instructional programme coherence can be identified in the Phillips et al. (2004) programme. They include a common instructional framework for teaching literacy across all schools involved in the programme; teachers working together to implement the common programme over a sustained period of time; and assessments which are common across time. Both the New Zealand programme and the high-programme-coherence schools in the USA rely on long-term partnerships between schools and external support organisations; the development of a common framework for literacy diagnosis which every teacher has to implement; expected collaboration between teachers; joint decision-making around assessments to use; and similar factors.

Underlying many of the features of schools with high programme coherence is the use of evidence to guide and evaluate teaching practices. For example, the aim of AUSAD was for practitioners to use student achievement data to inform practice. This has led directly to planning how to design classroom programmes that specifically meet the needs of students in these diverse urban schools. The research-schools partnership in this present study has responded to the increasing calls for greater understanding of the teaching and learning of comprehension to inform practice in New Zealand (e.g., Literacy Task Force, 1999; Learning Media, 2003) and internationally (Pressley, 2002).

Similarly, critical analysis of student data is identified as significant in school and teaching effectiveness research (e.g., Hawley & Valli, 1999; Robinson & Lai, 2006). In their literature review on effective professional development, Hawley and Valli (1999) identify critical analysis as a more effective form of professional development than traditional workshop models. The collection, analysis, and discussion of evidence were present in the schools maintaining gains in the Phillips et al. (2004) programme (Timperley et al., 2003).

A general question that arises is how much the critical analysis process contributes to the student changes in successful programmes. In the research and development programme reported here, the question concerns its contribution to the development of more effective teaching of reading

comprehension in schools serving culturally and linguistically diverse students in low-income communities. The collection, analysis, and discussion process took place in the context of collective analytic and problem-solving skills, where teachers collaborated with researchers and professional developers to co-construct the professional development. It is important to note here our assumption that professional expertise was distributed within and across schools, and that teachers would be able to contribute as coparticipants in a research-based collaboration (McNaughton, 2002). The issue of how teachers are viewed is particularly salient in the New Zealand context, as recent research syntheses show that school effects are consistently smaller than teacher or class-level effects. These latter effects can account for up to 60 percent of the variance in student achievement, depending on the subject areas, level of schooling, and outcome of interest, as estimated by Alton-Lee (2004).

This sort of collective problem solving represents one way of balancing two tensions identified in effective educational interventions (Coburn, 2003; Newman et al., 2001). One tension is around the issue of guaranteeing fidelity by adhering to a set of instructional procedures used in well-researched interventions, versus developing procedures which are derived from context-specific problem solving, but may have a less-well-known research-intervention base. A related tension is between importing a set of procedures, in a way which risks undermining local autonomy and efficacy, and a more collaborative development of common procedures, which risks losing instructional coherence. It seems to us that it is possible to construct fidelity to a common programme which has been strongly contextualised by developing a highly focused collaborative context. There is research evidence that suggests approaches in which professional development focuses on joint problem solving around agreed evidence, such as student achievement outcomes, is more likely than predetermined programmes to result in sustainable improvements in student achievement, particularly in reading comprehension (Coburn, 2003; Hawley & Valli, 1999; Timperley et al., 2003).

Evidence is critical to the processes of developing a professional learning community capable of solving the instructional problems associated with more effective teaching. Systematic assessment for formative and diagnostic purposes is essential in order to avoid the problems we have found before, where educators assume that children need a particular programme or approach, but close inspection of the children's profiles shows that they already have the skills targeted in those approaches (McNaughton, Phillips, & MacDonald, 2003). The significance of collecting and analysing data, rather than making assumptions about what children need (and what instruction should look like), was recently underscored by Buly and Valencia (2002). Policymakers in the State of Washington had mandated programmes without actually analysing profiles of low-progress students, identified by test scores from fourth grade National Assessment of Educational Progress (NAEP) scores. The assumption underlying policies and interventions was that poor performance reflected students' difficulties with more basic decoding abilities. Yet there was little data about this assumption, and little evidence to show that focusing on such skills would improve comprehension at fourth grade.

Using a broad band of measures, Buly and Valencia identified five groups of low-progress readers, some of whom did indeed have limited fluency and accuracy in decoding. However, mandating phonics instruction for all students who fell below the proficiency levels had missed the needs of the majority of students, whose decoding was strong, but who struggled with comprehension or language requirements for the tests. This finding highlights the need for research-based applications of best practice, based on analyses of student needs. One particular need that has been identified in other countries is for more effective teaching of reading comprehension than has typically been the case (Sweet & Snow, 2003).

The issue of sustainability

Developmental sustainability

A major challenge has been created by the advances made in schooling improvement and increasing instructional effectiveness through professional development. This is the issue of sustainability (Coburn, 2003). For the following research and development programme, sustainability has two meanings. The immediate concern facing the schools in South Auckland has been the need to build further progress in literacy, adding to the more effective instruction in the early years. This inevitably means considering the quality of the teaching and learning of comprehension (Sweet & Snow, 2003).

The issue in the decile 1 schools is that the subsequent instructional conditions set channels for further development, and if the channels are constructed for relatively "low" gradients of progress, this creates a need for further intervention. Unfortunately, as we have already noted and describe further below, the available evidence shows that despite the gains in decoding, there were still wide and possibly increasing disparities in achievement on comprehension tasks for Māori and Pasifika children, particularly in low-decile schools (Flockton & Crooks, 2001; Hattie, 2002; Lai et al., 2004).

The reason for needing to deliberately build this sustainability resides in the developmental relationships between decoding and comprehension. Logically, there are relationships such as the one identified by Tan and Nicholson (1997), who showed that poor decoding was associated with poor comprehension. It makes perfect sense that if you can't get the words off the page, you can't comprehend. The problem is that the corollary doesn't apply—decoding may be a necessary condition, but it is not a sufficient condition. So being a better decoder does not automatically make you a better comprehender.

The developmental reason for this can be found in Paris's (2005) multiple-components model of literacy development, or Whitehurst and Lonnigan's (2001) "inside outside" model of the strands of literacy development. Each of these explains that there are different developmental patterns associated with acquisition for components such as items, and for language meaning and uses, and

they are somewhat independent. This accounts for the phenomenon of rapid, accurate decoders who are not able to comprehend, which is described by professional educators and researchers (McNaughton et al., 2004). There is another developmental reason. Inoculation models do not apply to most phenomena in teaching and learning; just because you know and can do some stuff this year doesn't mean that you automatically make further gains next year. It depends at least in part on whether the teacher you meet effectively enables you to build on to and extend your learning. Fluent, accurate decoding is a necessary but not a sufficient condition for developing further comprehension skills (Block & Pressley, 2002; Sweet & Snow, 2003).

Sustainability of an effective professional learning community

There is a second meaning for sustainability. We now need to know which properties of teaching practices in schools enable success to be sustained with new cohorts of students and new groups of teachers joining schools (Timperley, 2003). Although effective practices may be able to be identified, this is an additional challenge. Sustaining high-quality intervention, it now seems, is dependent on the degree to which a professional learning community is able to develop (Coburn, 2003; Toole & Seashore, 2002). Such a community can effectively change teacher beliefs and practices (Annan, Lai, & Robinson, 2003; Hawley & Valli, 1999; Timperley & Robinson, 2001).

Several critical features of a collaboration between teachers and researchers are predicted to contribute to such a community developing (Coburn, 2003; Toole & Seashore, 2002; Robinson & Lai, 2006). One is the need for the community's shared ideas, beliefs, and goals to be theoretically rich. This shared knowledge is about the target domain (in this case, comprehension); but it also entails detailed understanding of the nature of teaching and learning related to that domain (Coburn, 2003). Yet a further area of belief that has emerged as very significant in the achievement of linguistically and culturally diverse students in general, and indigenous and minority children in particular, is the expectations that teachers have about children and their learning (Bishop, 2004; Delpit, 2003; Timperley, 2003).

Being theoretically rich requires consideration not only of researchers' theories, but also of practitioners' theories, and of adjudication between them. Robinson & Lai (2006) provide a framework by which different theories can be negotiated, using four standards of theory evaluation. These standards are accuracy (empirical claims about practice are well founded in evidence); effectiveness (theories meet the goals and values of those who hold them); coherence (competing theories from outside perspectives are considered); and improvability (theories and solutions can be adapted to meet changing needs, or to incorporate new goals, values, and contextual constraints).

This means that a second feature of an effective learning community, already identified above, is that their goals and practices for an intervention are based on evidence. That evidence should draw on close descriptions of children's learning as well as descriptions of patterns of teaching. Systematic data on both learning and teaching would need to be collected and analysed together. This assessment data would need to be broad based, in order to understand the children's patterns

of strengths and weaknesses, to provide a basis for informed decisions about teaching, and to clarify and test hypotheses about how to develop effective and sustainable practices (McNaughton et al., 2006). This means that the evidence needs to include information about instruction and teaching practices.

However, what is also crucial is the validity of the inferences drawn, or claims made, about that evidence (Robinson & Lai, 2006). The case reported in Buly & Valencia (2002), for example, shows how inappropriate inferences drawn from the data can result in interventions that are mismatched to students' learning needs. Robinson & Lai (2006) suggest that all inferences be treated as competing theories and evaluated.

So a further required feature is an analytic attitude to the collection and use of evidence. One part of this is that a research framework needs to be designed to show whether and how planned interventions do in fact impact on teaching and learning, enabling the community to know how effective interventions are in meeting its goals. The research framework adopted by the community needs therefore to be staged so that the effect of interventions can be determined. The design part of this is by no means simple, especially when considered in the context of recent debates about what counts as appropriate research evidence (McCall & Green, 2004; McNaughton & MacDonald, 2004).

Another part of the analytic attitude is critical reflection on practice, rather than a comfortable collaboration in which ideas are simply shared (Annan et al., 2003; Ball & Cohen, 1999; Toole & Seashore, 2002). Recent New Zealand research indicates that collaborations which incorporate critical reflection have been linked to improved student achievement (Phillips et al., 2004; Timperley, 2003) and to changed teacher perceptions (Timperley & Robinson, 2001).

A final feature is that the researchers' and teachers' ideas and practices need to be culturally located. We mean by this that the ideas and practices that are developed and tested need to entail an understanding of children's language and literacy practices, as these reflect children's local and global cultural identities. Importantly, this means knowing how these practices relate (or do not relate) to classroom practices (New London Group, 1996).

The main research project

This project is a result of a three-year research and development partnership between schools in the Otara: The Learning Community School Improvement Initiative, the initiative leaders and School Achievement Facilitators, the Woolf Fisher Research Centre at The University of Auckland, and Ministry of Education representatives. The representatives from an original group of eight schools formed an assessment team to work with researchers, the Ministry of Education and the initiative leaders on developing an intervention to raise student achievement.

The collaboration involved a replication of an innovative approach to research–practice partnerships. The purpose was to determine the extent of the challenges for effective teaching of comprehension, and to create better teaching methods to meet those challenges. As part of this, a cluster-wide intervention for all teachers teaching classes at Years 5–8 in the eight schools took place. This required extensive school-based professional development, as well as systematic collection of achievement data and classroom observations within a rigorous research design. The present research-based intervention was designed to test and replicate both the discrete components of effective teaching in school-wide implementation, and the model developed for a research–school practice partnership.

Research questions

This study: aims and research questions

This present study ("the Otara study") is a replication of a previously reported three-year research and development partnership to raise student achievement in reading comprehension ("the Mangere study"'). It aims to experimentally test hypotheses about how to raise the achievement in reading comprehension of students in a cluster of Otara schools, through a planned and sequenced research-based collaboration. The study addresses several areas of strategic importance to New Zealand, as previously noted.

The study also addresses specific theoretical questions. These are to do with the development of reading comprehension; effective instruction for reading comprehension; the development and role of professional learning communities; the role of (contextualised) evidence in planned interventions; and the nature of effective research collaborations with schools and the nature of replications. The specific research questions were:

- can the process and outcome of the previous intervention be replicated?
- can features of teaching and learning be identified for effective instructional activities that are able to be used by teachers to enhance the teaching of comprehension for Māori and Pasifika children in Years 5–8 in decile 1 schools in Otara?
- in what respect are they the same or different from a previous cluster of decile 1 schools in Mangere?
- can a research–practice collaboration develop a cluster-wide professional development programme that has a powerful educationally significant impact on Māori and Pasifika children's comprehension at Years 5–8 in decile 1 schools in Otara?

A general hypothesis derived from these areas is that instructional approaches to reading comprehension present in the cluster of schools could be fine-tuned to be more effective in enhancing achievement through a research–practice collaboration, and the development of professional learning communities, using contextualised evidence of teaching and learning.

The research base for each of these areas is outlined in the following sections.

What this report covers

This report describes the results of the research and development programme in action, as researchers and practitioners developed communities to meet the challenge of building more effective instruction for reading comprehension in linguistically and culturally diverse urban schools. The design methodology and frameworks for the interventions are described in Section Two. Section Three describes the results of these interventions for the overall three-year research and development partnership between schools and researchers. In the final section, results are summarised and discussed.

2. Methods

The overall partnership involved schools in the Ministry of Education, Otara: The Learning Community School Improvement Initiative, the initiative leaders, the Woolf Fisher Research Centre at The University of Auckland, and Ministry of Education representatives.

Main study participants

Schools

The Otara study originally involved eight decile 1 Otara schools. Six of these schools were contributing schools (Year 1–Year 6), one was an intermediate school (Year 7–Year 8), and one was a middle school (Year 7–Year 9). The schools ranged in size from 62 students to 470 students. One of the schools participated in only the first round of data collection; hence the analyses of gains are for seven schools, while the baseline profiles include eight schools.

Students

We report on several overlapping groups of students. The first group consists of all the students present at the beginning of the three-year study (Baseline sample). The second consists of one cohort of students who were followed longitudinally, starting from Year 4. The third group consists of all students who were present at the beginning and at the end of each year. The fourth group was all students present at any one time point.

Overall baseline samples

Baseline data using STAR and PAT (February 2004) were collected from 1646 students in eight schools. Different combinations of students sat the STAR and PAT tests for various reasons (e.g. being absent when the STAR test was administered but present when the PAT was administered). In addition, one school was unable to participate in the first round of STAR data collection.

The numbers of students at each year level were: Year 4 (mean age 8 years) n = 298; Year 5 (mean age 9 years) n = 311; Year 6 (mean age 10 years) n = 339; Year 7 (mean age 11 years) n = 370; and Year 8 (mean age 12 years) n = 328. The total group consisted of almost equal proportions of males and females (53% and 47% respectively) from 14 ethnic groups. Four main ethnic groups made up 93 percent of the sample. These groups were Samoan (37%), Māori (22%),

Cook Island (19%), and Tongan (15%). Approximately half the children had a home language other than English.

Longitudinal cohorts

One cohort of students was followed longitudinally from Time 1 to Time 6; these were those students who were Year 4 at Time 1 and who were present at all six time points, a total of 98 students. This cohort was labelled Cohort 1. (We could only follow one cohort of students over three years because schools in the sample were either contributing primaries (Years 1–6) or were intermediate and middle schools.) These students were a subset of the students included in the baseline sample.

Overall group year by year

A third group of students were those present at the beginning and end of each year. In Year 1 (Phase One, 2004) there were n = 973; in Year 2 (Phase Two, 2005) there were n = 924; and in Year 3 (Phase Three, 2006) there were n = 663. All of the students who were in the longitudinal cohort group were part of these groups, but these groups also included students who were present for only a single year, including Year 7 and Year 8 students, new Year 4 students (in the second and third year), new students arriving at the school and staying at least a year, and students who were present for a year only. These data do not include one school that pulled out after participating in the baseline sample. Phase Three does not include one school which pulled out after Phase Two.

Total school population

A fourth group of students were all the students present at each time point and so included new students and students who subsequently left the school. The number varied from 1374 (Time 1) to 814 students (Time 6).

Teachers

Around 50 teachers were involved in each year of the project, including literacy leaders. Characteristics of the teachers varied somewhat from year to year, but in general around two-thirds had five or more years of experience, and 10 percent were beginning teachers. In the second year, 25 percent of the teachers were Pasifika or Māori.

School reading comprehension lessons

Observations were carried out early in the intervention (see below), and they provided a general description of the programmes across phases through which the intervention was delivered. Generally the programme was similar across classes and schools, and similar to the general descriptions of New Zealand teaching in the middle grades (Ministry of Education, 2006a; Smith & Elley, 1994).

The general structure of the reading programme comprised a daily session for 60 to 90 minutes starting with a whole-class activity, typically reading to children or shared reading, followed by small-group work. Children were grouped by achievement levels (using reading ages from assessments of decoding and comprehension) into between three and five groups. The teacher typically worked with one or two groups per day, usually using a guided-reading approach or a text/task study. They often used the same text for a week, which might be linked to topic study (in social studies or science) or a genre focus. Most classes observed (n = 10) had Sustained Silent Reading (SSR) for 10–15 minutes and forms of Buddy reading or Reciprocal Reading/teaching. When working independently, groups completed worksheets on text-related tasks such as identifying main themes or doing a text reconstruction, answering comprehension questions, preparing book reviews, or word study. In some classrooms, deliberate discussion, analysis, and practice with test formats occurred in which the skills required by tests such as cloze tests were highlighted.

A noticeable feature in classrooms was the use and explication of technical terms relating to written language and comprehension: for example, explicit identification and discussion of strategies (such as reciprocal reading, finding the main ideas, predicting), types of texts (such as story/article, genre); properties of types of texts (plot, setting, problem) or language unit types (such as synonyms, adjectives, "wh" questions, opinion, and even "contextual clues").

There was some variation in this generalised description. For example, with older classes or special timetabling arrangements, a lower frequency of class reading sessions occurred. Four of the classrooms varied in frequency of shared reading (conducted weekly), and in two classrooms the guided and shared reading session occurred once a week. A noticeable and typical feature of the lessons was the high levels of engagement which was maintained for up to ninety minutes, both by the teacher-led groups and the independent groups. Teacher aides were present in several classrooms, often working with the lowest-achievement group.

Design

Rationale for the quasiexperimental design

At the core of the following analyses is a quasiexperimental design from which qualified judgements about possible causal relationships are made. While it has been argued that the gold

standard for research into schooling improvement is a full experimental design, preferably involving randomised control and experimental groups over trials (McCall & Green, 2004), a quasiexperimental design was adopted for two major reasons. The first is the inapplicability of a randomised control-group design for the particular circumstances of this project. The second is the usefulness of the quasiexperimental design format, given the applied circumstances.

Schools are open and dynamic systems. Day to day events change the properties of teaching and learning and the conditions for teaching and learning effectively. For example, in any one year teachers come and go, principals may change, the formula for funding might be altered, and new curriculum resources can be created. More directly, teachers and schools constantly share ideas, participation in professional conferences and seminars adds to the shared information, and new teachers bring new knowledge and experiences. Such inherent features of schools are compounded when the unit of analysis might be a cluster of schools who deliberately share resources, ideas, and practices.

This "messiness" poses tensions in a randomised experimental and control-group design. On the one hand, the internal validity need is to control these sources of influence so that unknown effects do not eventuate which may bias or confound the demonstration of experimental effects. On the other hand, if schools are changed to reduce these influences so that, for example, there is no turnover in teaching staff, external validity is severely undermined because these conditions may now not be typical of schools in general.

It is of course possible to conceive of selecting sufficiently large numbers of teachers or schools to randomly assign. Then one assumes that the messiness is distributed randomly. If the teachers and the schools in the total set are "the same", then the error variance associated with this messiness is distributed evenly across experimental and control classrooms and schools. Leaving aside the challenges which large numbers of schools pose, a problem here is the assumption that we know what makes teachers and schools similar, and hence are able to be sure about the randomisation process. This is a questionable assumption to make. For example, in our previous study with a similar cluster of schools (McNaughton et al., 2006) the presence of bilingual classrooms in some schools, with different forms of bilingual provision, created difficulties for random assignment as well as for comparability across teachers, let alone across schools. So what counts as an appropriate control is not necessarily known. There may also be insufficient instances of different types of classrooms or schools even to attempt random assignment.

There is another difficulty: that of withholding treatment from the control group of schools. Just about any well-resourced, planned intervention is likely to have an effect in education (Hattie, 1999). The act of deliberately withholding treatment, as required in control-group designs, raises ethical concerns. Some researcher groups in the United States, also concerned with educational enhancement for schools serving poor and diverse communities, have deliberately adopted alternatives to randomised experimental and control-group designs because of ethical concerns for those settings not gaining access to the intervention (Pogrow, 1998; Taylor et al., 2001). Hattie (1999) proposed that the ethical difficulty could be overcome by comparing different

interventions, thus not withholding potential benefits from any group. This is not always a workable solution, for example when the theoretical question is about the effects of a complex, multicomponent intervention that reformats existing teaching in a curriculum area, such as literacy instruction. Here there is no appropriate alternative intervention other than existing conditions. The American Psychological Association has detailed guidelines for conditions under which withholding treatment is justified. For example, if an intervention is shown to be effective, then it should be implemented in the control group. This route has similarities with the design proposed below.

The most problematic aspect however, is the underlying logic of experimental and control-group designs. In these designs, the variation within each group (given the simple case of an experimental and a control group) is conceived as error variance and, when substantially present, is seen as problematic. The alternative design adopted below is based on a view of variability as inherent to human behaviour generally (see Sidman, 1960), and specifically present in applied settings (Risley & Wolf, 1973). It deliberately incorporates variability and the sources of that variability into the design. Questions about the characteristics and sources of variability are central to knowing about effective teaching and learning, and can be explored within the design. Such a design is more appropriate to the circumstances of building effectiveness over a period of time, given that the variability is an important property (Raudenbusch, 2005). Similarly, such designs are useful in the case of planning for sustainability with ongoing partnerships. In fact, longitudinal designs are recommended in which sources of variability are closely monitored and related to achievement data, such as levels of implementation, the establishment of professional learning communities, coherence of programme adherence, and consistency of leadership and programme focus over time (Coburn, 2003). These are all matters of concern in the research reported here.

Repeated measures of children's achievement were collected in February 2004 (Time 1), November 2004 (Time 2), February 2005 (Time 3), November 2005 (Time 4), February 2006 (Time 5), and November 2006 (Time 6) as part of the quasiexperimental design (Phillips et al., 2004). The design uses single case logic within a developmental framework of cross-sectional and longitudinal data. The measures at Time 1 generated a cross section of achievement across year levels (Years 4–8), which provided a baseline forecast of what the expected trajectory of development would be if planned interventions had not occurred (Risley & Wolf, 1973). Successive stages of the intervention could then be compared with the baseline forecast. The first of these planned interventions was the analysis and discussion of data. The second was the development of instructional practices. The third was a phase in which sustainability was promoted. This design, which includes replication across cohorts, provides a high degree of both internal and external validity. The internal validity comes from the in-built testing of treatment effects described further below; the external validity comes from the systematic analysis across schools within the cluster.

The cross-sectional baseline was established at Time 1 (February 2004). Students from that initial cross section were then followed longitudinally and were retested at Times 2, 3, 4, 5, and 6,

providing repeated measures over three school years. Two sorts of general analyses using repeated measures are possible. Analyses can be conducted within each year. These are essentially before and after measures. But because they are able to be corrected for age through transformation into stanine scores (Elley, 2001), they provide an indicator of the impact of the three phases against national distributions at similar times of the school year. However, a more robust analysis of relationships with achievement is provided using the repeated measures within the quasiexperimental design format. They show change over repeated intervals.

As argued in the introduction chapter, good science requires replications (Sidman, 1960). In quasiexperimental research, the need to systematically replicate effects and processes is heightened because of the reduced experimental control gained with the design. In the design used with the present cluster of schools, there were inbuilt replications across age levels and across schools within the quasiexperimental design format. These provide a series of tests of possible causal relationships. However, there are possible competing explanations for the conclusions of the cluster-wide results that are difficult to counter with the quasiexperimental design. These are the well-known threats to internal validity, two of which are particularly threatening in the design adopted here.[2]

The first is that the immediate historical, cultural, and social context for these schools and this particular cluster meant that an unknown combination of factors unique to this cluster and these schools determined the outcomes. Technically, this is partly an issue of "ambiguous temporal precedence", and partly an issue of history and maturation effects (Shadish, Campbell, & Cook, 2002).

A second is that the students who are followed longitudinally and were continuously present over several data points were different in achievement terms from those students who were present only in the baseline, and subsequently left. It might be, for example, that the comparison groups contain students who were more transient and had lower achievement scores. Hence over time, as the cohort followed longitudinally is made up of just those students who are continuously and consistently at school, scores rise. Researchers such as Bruno & Isken (1996) report lower levels of achievement for some types of transient students. This is partly an issue of potential selection bias, and partly an issue of attrition (Shaddish et al., 2002). As the projected baseline is based on the assumption that the students at baseline are similar to the cohort students, having a lower projected baseline may result in finding large improvements due to the design of the study, rather than to any real effects.

There are three ways of adding to the robustness of the design, in addition to the inbuilt replications, which meet the major threats. The first is to use as a comparison a similar cluster of

---

[2] Other threats to internal validity, such as regression, testing, and instrumentation, are handled by other aspects of the methods. For example, all students in all achievement bands were in the cohorts; similarly repeated testing occurred, but with instruments that were designed for the interval of repetition and with alternative forms.

schools that received the intervention at a different time. The first study with Mangere has that function as it started a year before the intervention in the Otara cluster of schools. It was possible to identify and examine the baseline levels in the Otara cluster of schools after a year of intervention in the Mangere cluster, to check whether levels in the Otara cluster had changed significantly associated with the intervention starting in the Mangere cluster. If that happened the judgement of causality would be severely undermined. The Mangere and Otara clusters of schools were similar in geographical location (neighbouring suburbs); in type (all decile 1 schools); in number of schools (n = 7–8); in number of students (n = approximately 1400–1700); in ethnic and gender mix (almost equal proportions of males and females from about 14 ethnic groups, the major groups being Samoan, Māori, Cook Island, and Tongan); in starting levels of achievement; and in prior history of interventions. The Otara cluster was measured exactly one year after the baseline was established in the Mangere cluster, reported in Lai et al. (2006).

The two baselines are shown in Figure 1 and Figure 2. The comparison shows that before the intervention, the second cluster of schools had similar levels of achievement to the Mangere cluster of schools. This was so even though there was a delay of a year; and after an intervention had been in place for a year in the Mangere cluster of schools. As we reported earlier, achievement levels had risen in those schools in the year prior to the present study. This comparison adds weight to our previous conclusions by establishing that there was no general impact on similar neighbouring area decile 1 schools operating over the time period of the intervention in the Mangere cluster of schools.

Figure 1 **Baseline (at Time 1) student achievement by Year level for Cluster 1 (Mangere)**

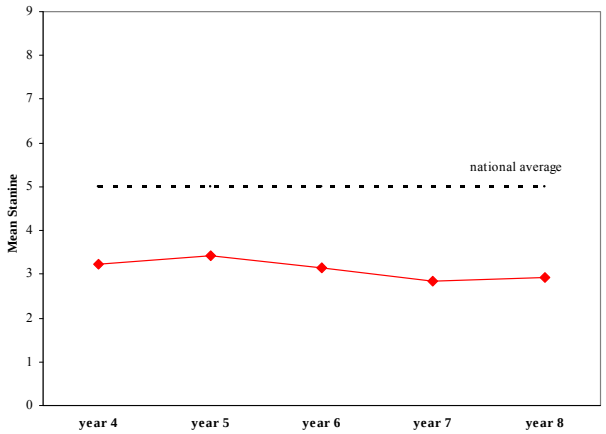


Figure 2 **Baseline (at Time 1) student achievement by Year level for Cluster 2 (Otara)**

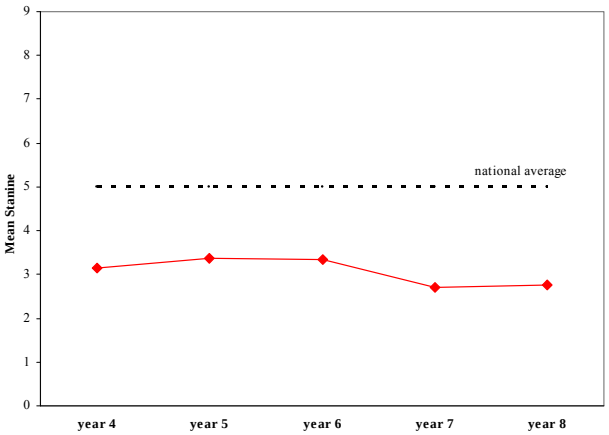
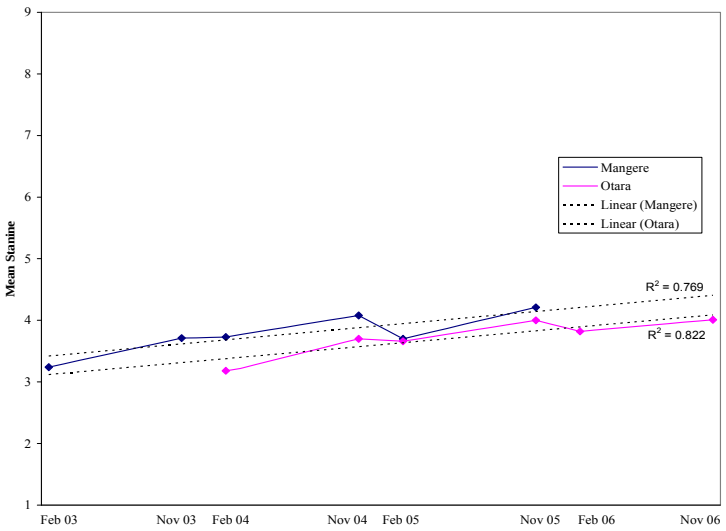


Figure 3 shows student achievement over the three years of the intervention in the Mangere cluster of schools, summarised by years, and the student achievement scores over the three years of the intervention in the Otara cluster of schools, summarised by years (these findings are repeated in detail below). What the figure shows is that the changes in student achievement did not happen at the same time. That is, in each cluster of schools the changes in student

achievement took place during the intervention and not before, thereby adding to the robustness of the design and indicating that the changes were unique to the intervention in the particular cluster of schools at that particular time.

Figure 3 **Cluster 1 (Mangere) and Cluster 2 (Otara) intervention summarised by years**



A second way of adding to the robustness of the design is by checking the characteristics of students who are included in the cross-sectional baseline analysis, but not included in the longitudinal analysis, because they were not present in subsequent repeated measures. This was done for the first year data in the Mangere and Otara clusters of schools by checking the Time 1 achievement data for those students who were present at two time points (Time 1 and Time 2), versus those students who were present only in the cross-sectional baseline established at Time 1. In the Otara project, the school which only participated in the baseline data collection was excluded from this analysis because the cross-sectional baseline needed to approximate the intervention schools as closely as possible). The results of this checking are given for the Mangere cluster of schools in Table 1 for raw scores, and in Table 2 for stanines and for the Otara cluster of schools in Table 3 for raw scores and in Table 4 for stanines. The comparisons indicate that unlike the Mangere cluster, in the case of the Otara schools, differences were detected at the baseline between those students who were present the whole year versus those present only at the beginning. The differences were at Years 4, 5, and 7. Because of this we have adjusted the quasiexperimental analyses comparisons. In the design-based comparison we only compare the longitudinal cohort with those 'baseline' students who were present for a whole year, thus providing a conservative baseline forecast.

Table 1 **Raw score means for Time 1 (Feb 03) by year level for Cluster 1 (Mangere)**

	Time 1 only (Feb 03) [a]			Time 1 pre-post (Feb 03) [b]				
	N	M	SD	N	M	SD	t value	ES
Year 4	34	16.26	7.57	205	16.9	6.82	0.50	0.09
Year 5	34	18.94	9.42	208	21.96	8.13	1.96	0.34
Year 6	30	23.60	9.58	265	24.09	8.87	0.28	0.05
Year 7	33	30.61	10.70	267	30.16	12.26	0.20	0.04
Year 8	34	32.68	13.18	271	37.41	13.11	1.99 *	0.36

*** P<0.001, ** P<0.01, * P<0.05

[a] This group contains those students who sat test at Time 1 only.
[b] This group contains those students who sat tests at both Time 1 and Time 2.

Table 2 **Stanine means for Time 1 (Feb 03) by year level for Cluster 1 (Mangere)**

	Time 1 only (Feb 03) a			Time 1 pre-post (Feb 03) b				
	N	M	SD	N	M	SD	t value	ES
Year 4	34	3.06	1.58	205	3.27	1.32	0.85	0.14
Year 5	34	2.88	1.81	208	3.52	1.52	2.22	0.38
Year 6	30	3.07	1.55	265	3.16	1.56	0.32	0.06
Year 7	33	2.85	1.15	267	2.84	1.31	0.56	0.01
Year 8	34	2.56	1.31	271	2.99	1.46	1.66 *	0.31

*** P<0.001, ** P<0.01, * P<0.05

[a] This group contains those students who sat test at Time 1 only.
[b] This group contains those students who sat tests at both Time 1 and Time 2.

Table 3 **Raw score means for Time 1 (Feb 04) by year level for Cluster 2 (Otara)**

	Time 1 only (Feb 04) a			Time 1 pre-post (Feb 04) b				
	N	M	SD	N	M	SD	t value	ES
Year 4	46	13.54	6.20	174	17.04	7.29	2.98 **	0.52
Year 5	30	18.33	9.44	217	21.92	8.08	2.23 *	0.41
Year 6	44	24.84	7.82	193	25.20	8.31	0.26	0.04
Year 7	29	25.79	10.10	216	30.19	10.56	2.12 *	0.43
Year 8	37	32.51	10.86	173	36.24	11.25	1.84	0.34

*** P<0.001, ** P<0.01, * P<0.05

[a] This group contains those students who sat test at Time 1 only.
[b] This group contains those students who sat tests at both Time 1 and Time 2.

Table 4 **Stanine means for Time 1 (Feb 04) by year level for Cluster 2 (Otara)**

	Time 1 only (Feb 04) a			Time 1 pre-post (Feb 04) b					
	N	M	SD	N	M	SD	t value		ES
Year 4	46	2.59	1.20	174	3.28	1.39	3.08	**	0.53
Year 5	30	2.87	1.61	217	3.46	1.51	1.99	*	0.38
Year 6	44	3.18	1.50	193	3.34	1.51	0.62		0.11
Year 7	29	2.28	1.10	216	2.78	1.22	2.12	*	0.43
Year 8	37	2.38	1.16	173	2.82	1.21	2.03	*	0.37

*** P<0.001, ** P<0.01, * P<0.05

[a] This group contains those students who sat test at Time 1 only.
[b] This group contains those students who sat tests at both Time 1 and Time 2.

These two additional checks add to the robustness of the design and should increase the believability of the results. Given significant outcomes, the design enables us to demonstrate that the intervention cannot easily be explained as arising from external and general effects on decile 1 schools in these suburbs, or the immediate histories of interventions and resourcing. In addition, they do not support the competing explanation that the students analysed in the longitudinal design were higher achievers anyway, and hence any "progress" is simply their usual levels compared with all other students.

A third way to add to the robustness is provided by analyses which are not part of the design logic. One already mentioned is in the form of pre- and post-testing, using normalised scores. In addition, and as an extension to these analyses, analyses are conducted of the overall student group at each testing time, irrespective of their previous presence or subsequent absence. This allows us to check whether overall achievement levels in schools could be influenced, despite new cohorts of students entering during the three years.

Procedures

Interventions across phases

Phase One—Analysis of data, feedback, and critical discussion

Current research on learning communities suggests that critical discussion and analysis of data may have an impact on effective practice that is independent of professional development programmes generally (Ball & Cohen, 1999; Timperley, 2003; Toole & Seashore, 2002). Theories about the needs for teaching and learning are developed through critical discussion, which is predicted to strengthen shared understanding and to inform current practices. In the current design

we are able to examine the effects of this process prior to the planned second phase of formal professional development.

Area-wide data were analysed by the school leaders and researchers in a meeting, then analysed by senior managers and senior teachers with each school using their specific school data. Additional sessions were conducted with support from one of the researchers (Mei Lai).

The analysis, feedback, and discussion process involved two key steps. Firstly, a close examination of student strengths and weaknesses and of current instruction practices to understand learning and teaching needs; and secondly, raising competing theories of the "problem" and evaluating the evidence for these competing theories. This meant using standards of accuracy, coherence, and improvability (Robinson & Lai, 2006). This process further ensured that the collaboration was a critical examination of practice and that valid inferences were drawn from the information. The feedback procedures with examples are described fully in Robinson and Lai (2006).

Phase Two—Targeted professional development

General outline

Targeted professional development, which took place in the second year, consisted of 10 sessions over two terms. It was designed using research-based examples and known dimensions of effective teaching. The sessions were led by one of the researchers (Stuart McNaughton). Five groups of 10–15 teachers with literacy leaders from different schools attended these half-day sessions, which occurred every two weeks from the middle of the first term of 2005. The last session was held at the end of the year. The curriculum for the sessions used a mixture of theoretical and research-based ideas, as well as teacher investigation and exemplification from their own classrooms.

Specific sessions

Session one introduced theoretical concepts of comprehension, and related these to the profiles of teaching and learning. A theoretical model was presented, drawing on Sweet and Snow (2003) and developmental analyses such as Whitehurst and Lonnigan (2001). A task was set to examine individual classroom profiles of achievement, and how these mirrored or differed from school and cluster patterns. Each session from this point started with group discussion of the task that had been set, and sharing of resources relating to the topic. Session two focused on strategies, in particular the issues of checking for meaning, fixing up threats to meaning, and strategy use in texts. A task to increase the instructional focus on checking and fixing was set. The third session introduced theories and research relating to the role of vocabulary in comprehension. Readings used included Biemiller (1999; 2001), Pressley, (2000), and others which identified features of effective teaching of vocabulary. The task for this session was to design a simple study, carried out in the classroom, which looked at building vocabulary through teaching. Sessions four and five identified the significance of the density of instruction and repeated practice, with a particular

focus on increasing access to rich texts, including electronic texts (Block & Pressley, 2002). The task mirrored this emphasis, with an analysis by the teacher of the range and types of books available in classrooms, and of engagement by different students. The sixth and seventh sessions introduced concepts of "incorporation" (of cultural and linguistic resources), and building students' awareness of the requirements of classroom tasks and features of reading comprehension (from McNaughton, 2002). Tasks relating to observing and analysing these features of instructions were set. Sessions eight and nine used transcripts of the video-recorded classroom lessons to exemplify patterns of effective teaching in different settings, such as guided reading and shared reading, and developed the practice of examining and critiquing each other's practices. The ninth session also had some specific topics which the groups had requested, such as the role of homework and teaching and learning in bilingual settings. Session nine also involved planning to create learning circles within schools, where colleagues observed in each other's classrooms aspects of teaching such as building vocabulary, and discussed what these observations indicated about effectiveness. The final session reviewed these collaborative teaching and learning observations.

Phase Three—Sustaining the intervention

The third phase was planned by the literacy leaders and researchers jointly. It involved four components. The collection, feedback, and critical discussion of achievement data continued. A second component was the continuation of the learning circles developed in the professional-development phase. A third was the development and use of planned inductions to introduce new staff to the focus on increasing reading comprehension and patterns of teaching and professional learning in the schools. The schools experienced staff turnover of differing degrees, but on average around a third of the staff changed from year to year. This component was designed to maintain and build on the focus with new staff. A fourth component was a teacher-led conference. School teams developed action-research projects, often with a pre- and post-testing component, to check various aspects of their programmes. The questions for these projects were generated by teams within schools. The researchers helped shape the questions and the processes for answering the questions through two group meetings with team leaders from each of six schools. Four of the schools developed research projects. Several of the research topics were about instructional strategies to increase vocabulary; others included evaluating the effectiveness of guided reading for reading comprehension, creating links between reading and writing; and three were more general projects looking at student achievement in the overall school programme. In each case, the projects involved use of formal or informal assessments of student outcomes. A total of 11 projects were presented in PowerPoint format at a conference after school hours in the fourth term of the school year, attended by 90 percent of the teachers involved. Other professional colleagues, such as literacy advisors, also attended the conference, and teachers involved in numeracy projects and leadership projects also presented.

Measures

Literacy measures in English

Initially baseline data on reading comprehension were collected, using both the revised Progressive Achievement Tests (PAT) in Reading (reading comprehension section only) (Reid & Elley, 1991), and the Supplementary Tests of Achievement in Reading (STAR) (Elley, 2001). The tests were what schools had decided to use as a group to measure reading comprehension, because they provided a recognised, standardised measure of reading comprehension which could be reliably compared across schools.

The revised PAT in Reading measures both factual and inferential comprehension of prose material in Years 4 to 9. Each prose passage consists of 100–300 words, and is followed by four or five multichoice options. The prose passages are narrative, expository, and descriptive, and different year levels complete different combinations of prose passages. The proportion of factual to inferential items per passage is approximately 50:50 in each year level.

STAR was designed to supplement the assessments that teachers make about students' "close" reading ability in Years 4 to 9 in New Zealand (Elley, 2001). The rationale behind STAR is the expectation that all students are to learn to read successfully at primary school. In other words, reading successfully at primary school means learning to read appropriate text fluently, independently, and with comprehension. The definition, according to the Literacy Task Force report (1999), implies that students should also be equipped with reading skills thought to be central to reading programmes at each level of the primary school, although some of them (e.g., critical reading, gathering information) may be given greater emphasis at the upper levels (Elley, 2001).

Analyses over the first year revealed that the correlation between the two tests was .54 ($P < .01$). In the test manual, Elley (2001) reported correlations between 0.70 and 0.78 for Year 4 to 8 students. This, as Elley suggests, indicates that the tests measure similar but not identical facets of reading comprehension. Subsequently, schools focused on using the STAR data, as PAT was used primarily to develop baseline profiles. The outcome data on reading comprehension reported here for the overall project are from across the six time points using STAR (Elley, 2001). These tests were designed for repeated measurement within and across years, are used by schools, and provide a recognised, standardised measure of reading comprehension which can be reliably compared across schools. In addition to these assessments, the schools used other reading measures for both diagnostic and summative purposes.

STAR subtests—Years 4–6

In Years 4–6, the STAR test consists of four subtests measuring word recognition (decode familiar words through identifying a word from a set of words that describe a familiar picture); sentence comprehension (complete sentences by selecting appropriate words); paragraph

comprehension (replace words which have been deleted from the text in a cloze format); and vocabulary range (find a simile for an underlined word). Only the paragraph comprehension subtest is not multichoice and consists of 20 items, 10 more items than the rest of the subtests. In Years 7–8, students complete two more subtests in addition to the four subtests described above: the language of advertising (identify emotive words from a series of sentences); and reading different genres or styles of writing (select phrases in paragraphs of different genres which best fits the purpose and style of the writer). In Years 7–8, there are 12 items per subtest, except for paragraph comprehension, which consists of 20 items. Both tests have high reliability and validity (Elley, 2001; Reid & Elley, 1991).

Subtest 1: Word recognition

Word recognition assesses how well students can "decode words that are familiar in their spoken vocabulary" (e.g., umbrella, dinosaur, cemetery). The test measures word recognition in the form of decoding of familiar words, through identification of a word from a set of words that describe a familiar picture. Ten pictures that are assumed to be familiar to students are in the subtest. Each picture has four words alongside it, one of which is correct. Students are asked to select the correct word that matches the picture. The words tested in STAR Test 4–6 are taken from Levels 6–8 of the NZCER Noun Frequency List (Elley & Croft, 1989), and are thus well within the range of most pupils' spoken vocabulary. Evidence shows that, for the majority of pupils in the upper levels of the primary school in New Zealand, word recognition is a skill that has been well mastered, but many schools have a few pupils who will struggle with this task.

Subtest 2: Sentence comprehension

Sentence comprehension assesses how well students can read for meaning. The prerequisite for this subtest is that students are able to read a range of very short texts (sentences) well enough to complete them with an appropriate word. Students are to complete the 10 sentences by choosing, from four words, the word that best suits the sentence. This test assesses decoding skills, and the ability to use a range of sources to gain meaning. To some extent, it also reflects students' mastery of the concepts of print, their vocabulary, and their ability to predict.

Subtest 3: Paragraph comprehension

Paragraph comprehension assesses students' reading comprehension by requiring them to replace words which have been deleted from the text (cloze format). Using the context of the text as cues to meaning, students can find it easier to replace the missing words, given that they can comprehend the text. The subtest shows how well pupils can apply the skills tested in subtest 2 to longer texts, when more linguistic and knowledge cues can be called on from previous sentences. Unlike subtests 1 and 2, this test is not multichoice. It does, however, consist of 20 items, 10 more than the other subtests. The students are required to fill in 20 blanks in three short paragraphs of

prose (paragraph 1 = 6; 2 = 7; 3 = 7), using the context of the surrounding text as cues for meaning.

Subtest 4: Vocabulary range

The development of a good reading vocabulary is the main focus of this test, because it measures students' knowledge of word meanings in context. Ten complete sentences are listed. One word in each sentence is in bold print and underlined. The students are required to circle one word from the four words under the sentence that means the same, or nearly the same, and is therefore close in meaning to the bold underlined word. The words included in this test are all taken from the New Zealand Oxford Primary School Dictionary of 30,000 words, and were selected after extensive trials had shown them to be of appropriate difficulty for the students in the relevant year groups.

STAR subtests—Years 7–9

In Years 7–9, students complete two subtests in addition to the four subtests described above. These subtests are on the language of advertising (identify emotive words from a series of sentences) and reading different genres or styles of writing (select phrases in paragraphs of different genres which best fit the purpose and style of the writer). In Years 7–9, there are 12 items per subtest, except for paragraph comprehension, which consists of 20 items. Both tests have high reliability and validity (Elley, 2001; Reid & Elley, 1991).

Subtest 5: The language of advertising

This subtest requires the students to identify emotive words which are typically used by advertisers when trying to attract consumers to buy. Students read a series of sentences and circle the one word that sounds appealing, but provides no information, e.g., "fabulous", "gotta-go", "cosy". This skill is part of learning to be a critical reader, and is stressed in "English in the New Zealand Curriculum" for Years 7 and 8 (or Curriculum Levels 4 and 5).

Subtest 6: Reading different genres or styles of writing

Pupils in the senior levels of primary school are expected to read with understanding various styles or genres of writing, both formal and informal. To assess this skill, pupils are given some paragraphs which represent a range of genres, and at particular points in each paragraph, they are asked to select the phrases which best fit the style and purpose of the writer. The genres represented include traditional fairy tales, business letters, informal letters, recipes, and computer manuals.

Reliability of STAR and PAT assessments

At the beginning of 2004, the Otara: The Learning Community (O:TLC) lead teachers developed an intraschool standardised process of administering the tests and moderating the accuracy of teacher scoring. This involved standardising the week and time (morning) of testing, and creating a system of randomly checking a sample of teachers' marking for accuracy of scoring. Accuracy of scoring was further checked by the data-entry team from Woolf Fisher Research Centre during data entry and during analysis. The STAR and PAT were administered as part of the schools' normal assessment cycle at the beginning of the school year, and thereafter for STAR at the end of each year also (using the parallel form). At Time 1 (February 2004), a number of additional analyses took place. These involved analysing student scores on factual and inferential questions in the PAT.

Observations

Evidence about current classroom instruction came from systematic classroom observations carried out by the researchers. These were designed to provide a sample of how features of teaching and learning might map on to the achievement data. Our argument was that a fully effective teaching analysis needed to examine classroom instruction, otherwise assumptions about what was or was not being taught would be unchecked.

Observations at baseline (first year)

Phase One observations were carried out in 15 classrooms in seven schools from Years 4–5 through to Year 8 (including two bilingual Samoan classrooms), selected to represent all schools and age levels within schools. In the first term, classroom instruction was observed during the usually scheduled core reading session within which the teaching and learning of reading comprehension occurred (average 46.07 minutes, SD = 14.83 minutes). Class sizes generally ranged from 21 to 26. Unlike the Mangere study, which used a combination of tape recording and in situ running records, the classroom reading sessions were video taped and later transcribed. Discussions with the teachers were held again, and the observers also made notes on features of the general classroom programme.

Observations at Time 3 and Time 4 (second year)

At the beginning of the second year (February 2005), six classrooms in seven schools from Years 4–5 through to Year 8 (including one bilingual Samoan classroom) were systematically observed; again, classrooms were nominated and selected to represent age levels. Video recordings were made of teacher–student and text interactions during the usually scheduled core reading session for comprehension. All of the whole-group and at least one small-group activity were recorded. The amount of time allocated to reading on a typical day ranged from 30 to 60 minutes (average

39.0 minutes, SD = 12.96 minutes). The observations were repeated at the end of the year (November) with seven teachers, only four of whom were video taped at the beginning of the year, and were still teaching and available for a second video taping. Discussions with the teachers were held again, and the observers also made notes on features of the general classroom programme.

Coding and reliability of observations

Systematic classroom observations of core literacy activities were coded, using the coding handbook developed by the research team from the first intervention study. The following explains the categories exemplified with transcript examples used in developing the coding categories.

Exchanges were the basic unit of analysis of classroom lessons transcribed from the videotapes. An exchange was initiated by an utterance followed by a set of interactions on the same topic, involving comments, questions, directions, explanations, or feedback between teacher and child or child and group. A minimal exchange would be one turn by a teacher or student. A change in topic during interactions, a return to text reading, or a new activity signalled the end of an exchange. Each exchange was coded as a specific type, using the following definitions:

- text related
- nontext related
- vocabulary elaboration
- extended talk
- checking and evaluating
- incorporation
- awareness
- feedback.

Text-related exchanges and nontext-related exchanges

Text-related exchanges were exchanges that dealt specifically with the text at hand. Any comment or question related to the text came under this category. Nontext-related exchanges, on the other hand, were exchanges that were not related to the text, but were employed nevertheless to prompt students to answer comprehension questions that were otherwise difficult to respond to. An example of text-related exchange:

> T: OK what do you think it means by [Our children have eyes but they cannot see]? Try and relate it back to his idea that he had. [repeat] What do you think it means by that statement?
>
> C: They're blind.
>
> T: They're blind. OK anything else?
>
> C: They have eyes, but they're not colour blue, green, or black. They're colour blind.

T: Alright any other ideas? Let's have a read. Just look at the sentence. Let's start from that. [The Sherpa thought for a moment] OK. This time I'll read it to you. [repeats. What we need more than anything is a school.] So what do you think he is trying to say? [repeats]

C: They can't learn properly.

T: They can't learn properly. Why S?

C: Because there's no school.

T: Because there's no school. They have eyes but how can they learn about the rest of the world if they're not taught? So Sir Edmund Hillary's ideas was to what?

C: Build a school.

T: Build a school so …

C: They can learn.

T: So they can learn. And so they can see all the different things that are happening in the world.

OK.

Vocabulary elaboration

Exchanges that elaborated vocabulary in the text were coded into three (non-mutually exclusive) types. These were questions seeking elaboration, elaboration of vocabulary by the teacher and elaboration by looking up meanings in the dictionary. An example of vocabulary-elaboration question and teacher comment:

Text word: *gesture*

T: [beckons'] Right so, we were stuck there on what it means … some people say what?

Gp: Gesture, to summon or to gesture.

T: Any other word that you can replace that word?

C: Summon or gesture.

T: Well, can I use the word 'invites'?

Extended talk

Extended talk meant conversations sustained over several turns on a topic that allowed the teacher or child to develop further features of place, time, theme, and concept. Exchanges that were limited to a synonym or brief comment on a word or phrase were not coded as extended talk. There were two types: extended talk by teacher, and extended talk by child or students. An example of extended talk:

T: What do you think is happening? What do you think the people are in the air?

C: Evil wizards?

T: You think the ones in the air are evil wizards?

C: People who are controlling the beast.

T: You still think it's the beast, yeah could be the beast.

C: The ones controlling the horse.

T: The Bulgarian people?

C: The dragon?

T: The four little people in the air is a dragon?

Checking and evaluating

These are exchanges in which there is some explicit reference to checking and evaluating evidence. The reference could involve questions, directions, prompts, feedback, or comments. It can be initiated by the teacher, a child, or a group and involve the teacher, the child, or a group. Three subcategories were noted. Teacher checking is where the teacher makes reference to students to check the accuracy of their responses by going back to the text to search for confirmation. Child checking is where the child checks the validity of the responses by verbalising what is found after the teacher prompts. The final subcategory involves the teacher and child checking for the evidence together. An example of checking and evaluating:

T: A kid … do we know exactly how old exactly it is S …?

C: Older like when you see some goats born and you remember their birthday um that's when you know how old they are.

T: Ok Ok so perhaps it's not a year old. Perhaps it's a month old.

C: We don't know Miss cos it doesn't say in the book.

Incorporation

Incorporation exchanges were those in which students' knowledge, skills, and other expertise are brought into an activity. It is a deliberate attempt by the teacher to make direct links between the text being read and the experiences of the student, through frequency of overt connection with topic events, and concepts that are familiar to the child—for example, prompting a child to talk about feelings, referring to past events or activities, or involving a child in the story. Furthermore, incorporation is also when the language of the child is used. An example of incorporation:

T: Put up your hands if you've been trapped by something before.

C: My hand got stuck in a machine, the eftpos machine.

T: Your ..?

C: My hand got stuck in the machine cos …

T: What kind of machine was it?

C: You had to put the money in.

T: You had to put the money in. And your hand got stuck?

C: Yeah.

T: My goodness. Did it take a long time for your mum and them to get your hand out?

C: I got it out myself.

T: You got it out by yourself. So what actually trapped your arm inside? Did it have a cover or a door or something on the machine?

C: Yes it had this little thing in front of it.

T: What was the … what little thing?

C: It was a kind of a door.

Awareness

Awareness means exchanges that focus on the child's awareness through teacher comments, questions, explanations, or feedback that explicitly draw attention to the relevance of the child's knowledge or reflection on knowledge, to the rules of participating, and to the purpose or ways of participating. The two types were awareness of strategy, such as clarifying, predicting, and summarising, and awareness of any other aspect of the task or child's expertise. An example of awareness, using context:

T: Context. Words that are related to it. It may be just a simple sentence but one word may make sense to you. Now that we have finished our lesson you're going to do (from OHP) [What new thing did you learn today and] …?

C: (all): [What more do I need to learn in order to achieve greater success?]

T: OK. That's success isn't that so?

C: Yes.

An example of awareness of another aspect:

T: OK students, we're going to do shared reading with me and our learning intentions, as you always say, is … I want you to understand what is "fact" and "opinion". That is the main focus … how to extract or identify a fact and a …

C: (all): Opinion.

T: Opinion. So we're going to do that, to work on exploring language that is actually picking adjectives from the passage and we will discuss the meaning of the words. We will

do some critical thinking. Also, we're doing a cloze activity … Right, so that is our learning intention of today's lesson.

Feedback

Feedback is defined as teacher responses reliant on a student action or verbal contribution. There were two subcategories of feedback: high feedback and low feedback. High feedback is any feedback that clarifies, elaborates on, or adds to the student's statement or response. It includes teacher correction of a student's incorrect answer or statement, or teacher response to a student's utterance with a question. Low feedback is nondescriptive and provides no extra information for the student, other than correctness. Examples of high feedback are given above in the example of extended talk and checking.

As reported in the Mangere study, development of the definitions and codes took place over several sessions in which three transcripts were randomly selected and coded by different members of the research team, until there was close to 100 percent agreement on the basic unit and on types of exchanges (all members of the team had to concur on presence or absence for an agreement to be scored). Subsequently, a further transcript was coded by each member independently, and the interobserver agreement calculated by the presence and absence of types of exchanges. The levels of interobserver agreement ranged from 86 percent for the awareness categories to 100 percent on the text-related or nontext-related categories. One member of the team coded all observational data used for the analyses presented here.

Data analysis

Reading comprehension achievement

The data were analysed in terms of patterns of achievement, using repeated measures and gain scores, as well as raw-score shifts. In addition to the use of raw scores on subtests and stanines, the analysis of achievement patterns for the STAR assessments used distribution bands. The manual (Elley, 2001) groups stanines into five bands; stanine 9 (outstanding, 4 percent of students); stanine 7–8 (above average, 19 percent of students); stanine 4–6 (average, 54 percent of students); stanine 2–3 (below average, 19 percent of students; stanine 1 (low, 4 percent of students). These bands were used to judge educational significance. SPSS (Statistical Package for the Social Sciences) and Excel programmes were used to create a database where data from all testing periods could be recorded and analysed.

Testing the effectiveness of the interventions was explored in a number of steps, using tests of statistical and educational significance. We compared means and distributions for the children in terms of both pre- and post-testing, and in terms of comparisons using the projected baselines. These comparisons use standard statistical procedures, such as t tests. A further step was

introduced to determine the educational significance of the interventions. This was based on an assessment of the effect size of the educational intervention. Effect size (ES) is a name given to a family of indices that measure the magnitude of a treatment effect. Hattie (1999) describes a 1.0 effect size as an increase of one standard deviation, which usually represents advancing student achievement by about one year. To measure the magnitude of a treatment of effect in this study, Effect Size Cohen's D was employed (Cohen, 1988).

Instruction

Two levels of analysis were carried out from the classroom observations. The first level was analysing changes in overall achievement and the text components from the pre- and post-testing. A second level involved both quantitative and qualitative analyses of two selected case studies of teachers from the second year. This enabled us to go beyond the frequency counts, using the transcript data and classroom records, to better understand the relationships with achievement and variability in achievement across classrooms.

3. Results

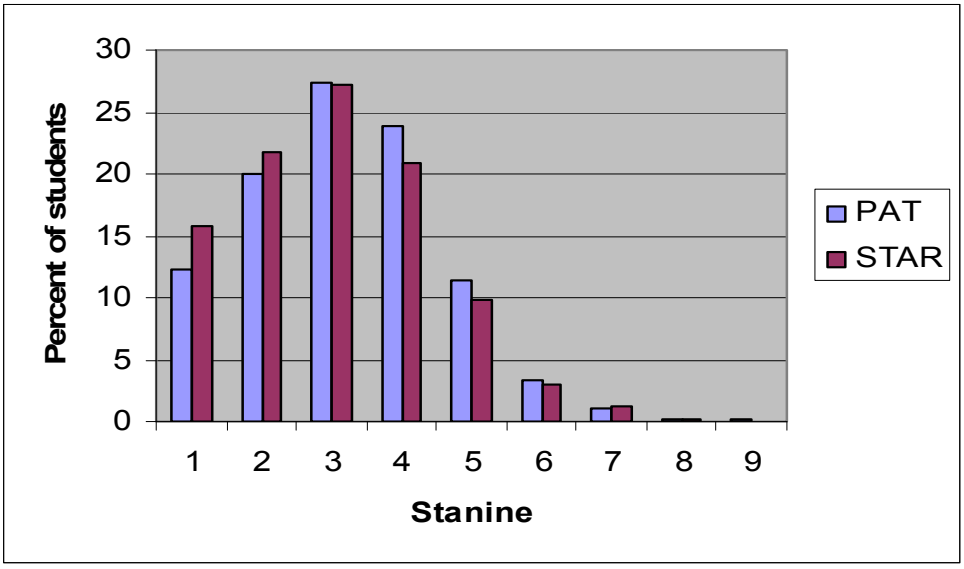
Baseline profile

The baseline (Time 1) results describe students who sat STAR or PAT at the beginning of the intervention (February 2004). This is presented in four sections: The general profile of reading comprehension, content analysis of PAT and STAR, and gender and ethnic-group breakdowns.

General profile of reading comprehension

The stanine distributions of both tests indicate that the average student experienced difficulty on these measures of reading comprehension. Figure 4 shows the stanine distribution in both tests across all year levels. The average student in both tests scored in the "below average" (stanine two and three) band of achievement. For PAT and STAR the mean stanine respectively was: 3.19 (SD = 1.42) and 3.02 (SD = 1.42). Both were in the "below average" band. Over 60 percent of students scored in the "low" (stanine 1) or "below average" (stanines 2 and 3) bands, about 30 percent were in the "average" band (stanines 4 to 6), and less than 2 percent in the "above average" or "superior" bands.

Figure 4 **Stanine distribution for PAT and STAR for year levels 4–8**



Across the year levels, the pattern was virtually the same in both tests, with the median in every year level at stanine three except in Year 5 for PAT (See Figures 5 and 6.) The relatively flat line in stanines across year levels indicates that under initial instructional conditions, children made about a year's gain for a year at school, remaining at two stanines below national average across years.

Figure 5 **Stanine distributions for PAT in Years 4–8**

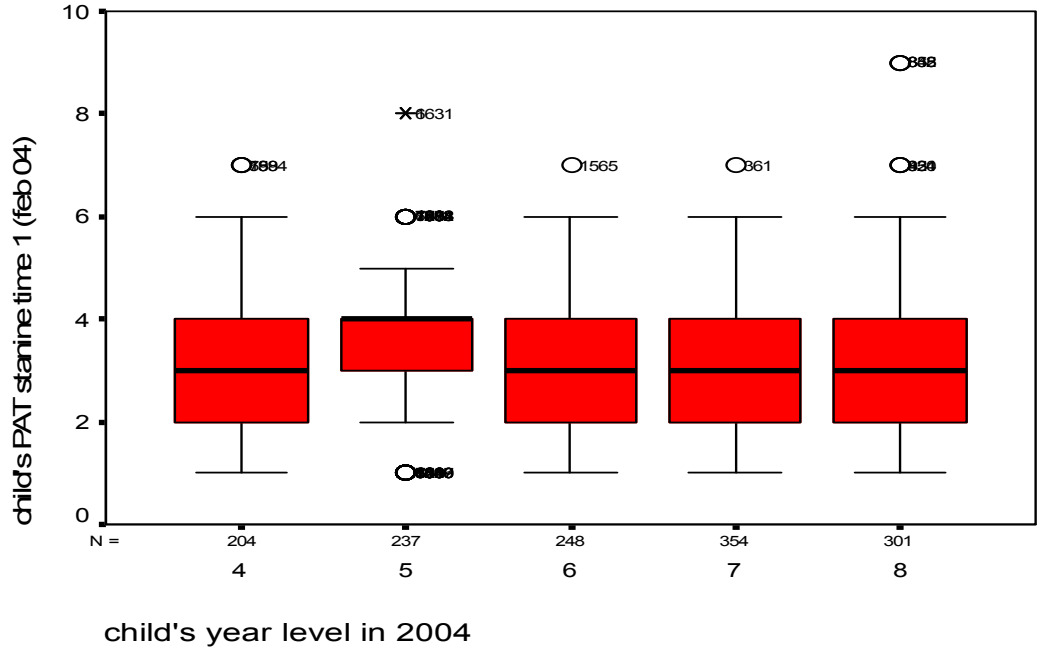
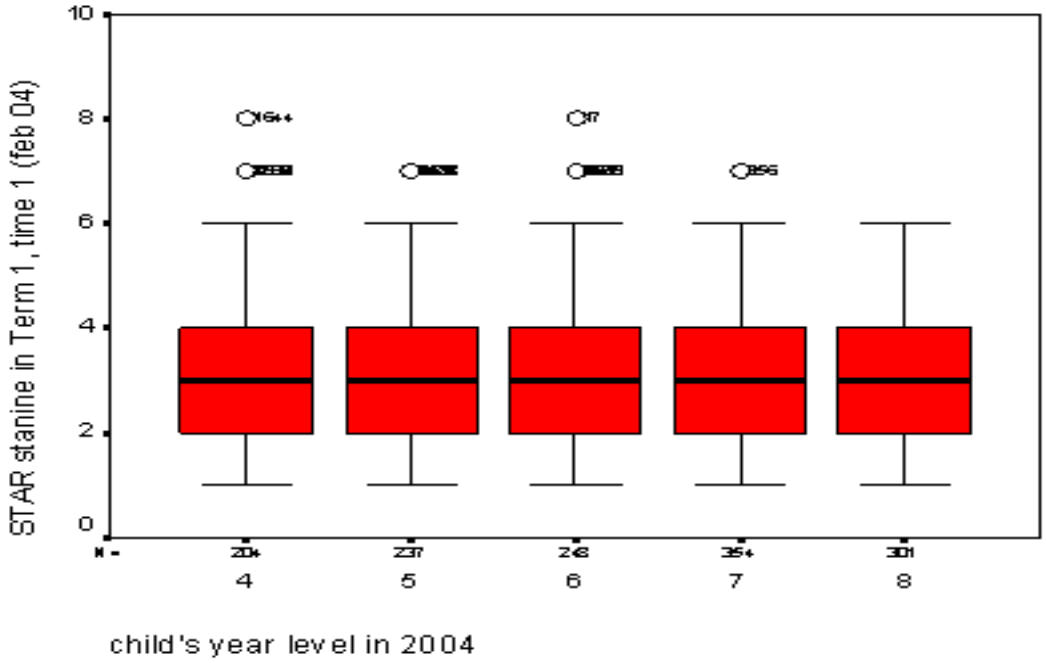


Figure 6 **Stanine distributions for STAR in Years 4–8**



Content analysis on the PAT

Apart from Year 7, mean scores on inferential questions in the PAT were lower than mean scores for factual questions (see Table 5). This pattern across the year levels suggests that students of different year levels experienced similar difficulties in answering both items but found the inferential items harder. This pattern is different from what we found in Mangere, but is what is expected. The developer of the test, Elley (personal communication, October 22, 2004), notes that the pattern of achievement for the two types of questions would be expected to be very similar in a large sample, but that factual questions are consistently easier than inferential questions. (Note that the maximum raw score for both factual items and inferential items was approximately 20; Reid & Elley, 1991). As expected, there was a significant correlation between factual and inferential items (r = .53, p < 000.) The pattern in these schools was different from the pattern in the Mangere cluster where the scores at each age level were very close.

Table 5 **Means (and standard deviations) of factual and inferential questions across year levels**

Year Level	N	Factual	Inferential
4	276	6.57 (3.53)	4.36 (2.52)
5	304	6.61 (3.20)	4.53 (2.47)
5	327	6.30 (3.44)	4.35 (2.68)
7	361	5.68 (2.99)	5.66 (3.01)
8	317	7.11 (3.82)	6.24 (3.07)
Total	1585	6.42 (3.43)	5.06 (2.88)

Content analysis on the STAR subtests

Analysis of the STAR subtests revealed consistent patterns across the subtests at each year level. Figures 7 and 8 show the average percentage obtained in each subtest. At every year level, students scored highest on subtest 1 (word recognition) and lowest on subtest 3 (paragraph comprehension), indicating that students in all year levels experienced more success in decoding words than comprehending a paragraph. A series of paired t tests between subtests averaged across years revealed that the means for subtest 1 were significantly higher than the means for the other subtests (t values all above t = 18.0, $p < .000$) and vocabulary was significantly lower than sentence (t = 6.86, $p < .000$) and paragraph (t = 16.86, $p < .000$) subtests (subtests 2 and 3). It should be noted that there are 20 items for the paragraph subtest whereas the others have only 10 items. All the subtests of STAR were significantly correlated ($p < .000$).

This pattern (including the average scores) was virtually identical to the pattern of scores in the previous cluster of South Auckland schools (Lai et al., 2004). They were also similar to the norms in the STAR manual, albeit lower. This suggests that students in Otara had a similar profile of strengths and weaknesses in reading comprehension to that of students in a similar geographic and

decile area, and students in New Zealand in general; albeit they experience more difficulties in each of the areas than typical students in New Zealand.

Figure 7 **Average percentages obtained in each subtest (STAR) for year levels 4–6**

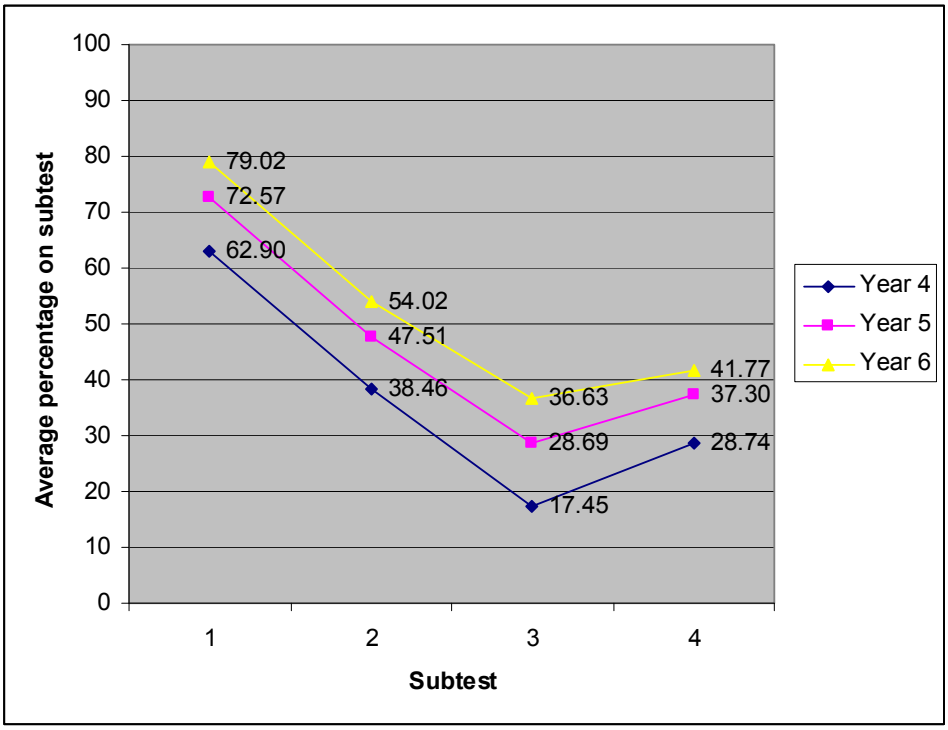
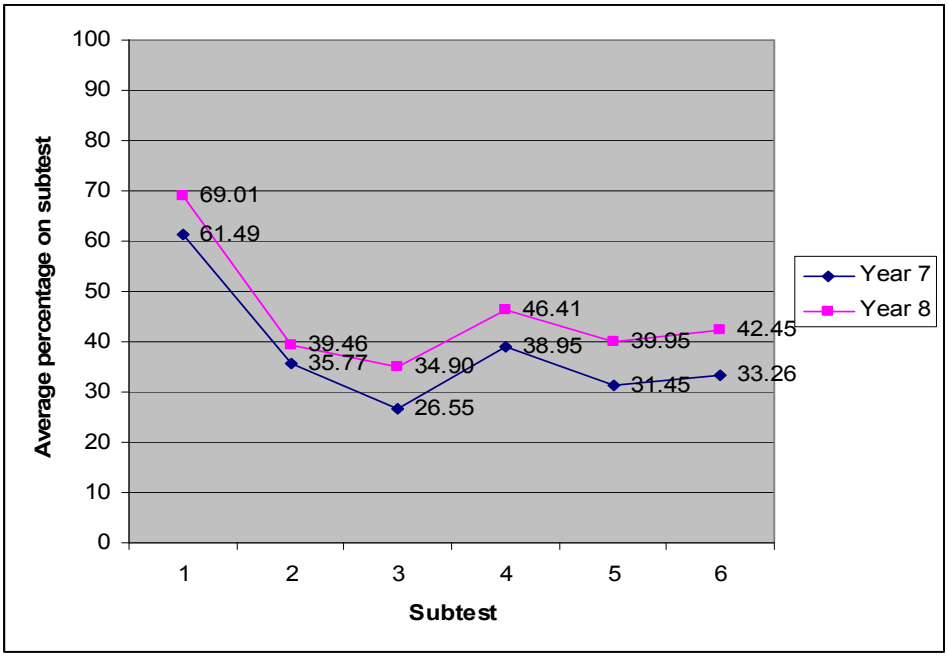


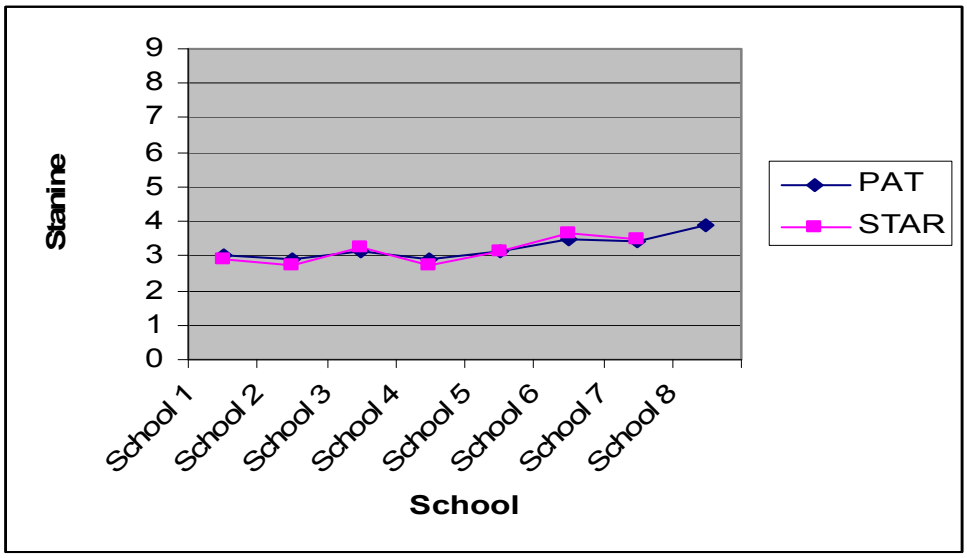
Figure 8 **Average percentages obtained in each subtest (STAR) for year levels 7–8**



School profiles

The average student in all schools experienced difficulty on both measures of reading comprehension. The average student in every school scored in the "below average" (stanine two and three) band of achievement, albeit one school's average was only 0.1 of a stanine below the average band of stanine 4–6 (see Figure 9). (Note that one school did not complete baseline information for STAR.) But the variability between schools suggests the significance of school-related factors such as quality teaching.

Figure 9 **Mean scores for schools for PAT and STAR**



Ethnicity and gender

Analyses were also conducted by ethnicity and gender. For the PAT the pattern of achievement was consistent across the major ethnic groups. The median stanine was 3, with 50 percent of students scoring between stanine 2 and 4. In the STAR test, however, the pattern of scores was slightly different between NZ European students and the students from the other ethnic groups. Whilst the median for NZ European students was the same, 50 percent of students scored between stanine 2 and 5. The other ethnic groups all scored a median of 3 with 50 percent scoring between stanine 2 and 4 in STAR. Similarly, there were few gender differences between males and females in both PAT and STAR. The median for both males and females was stanine 3, with 50 percent of students falling between stanine 2 and 4. These results suggest that students from the main ethnic groups, and males and females, experienced similar difficulties in reading comprehension.

Classroom instruction profile

At the beginning of the research programme, classroom instruction was viewed as an open problem, the critical aspects of which needed to be understood for this context. General principles of comprehension development and learning on the one hand, and instruction on the other hand,

informed what was to be observed. The coding system developed in the Mangere study was employed in the first year, assuming that there were likely to be similar issues in teaching and learning given the close similarities between the communities of teachers, children, and families, but qualitative analyses of transcripts from the 15 teachers were also used, and specific aspects of the resulting profile were developed in situ, in the form of hypotheses. In essence the approach had a relatively open-ended investigative purpose but was theory driven. In addition to the coding system and the qualitative analyses, an additional analysis was completed looking at vocabulary instruction. Every instance where the teacher explicitly defined, described, or elaborated a word during the lesson was examined and checked to note the type of word. Three categories of words were identified: topic-related words, technical language words, and other words.

The resulting instructional focus can be summarised into the same areas as we developed for the Mangere cluster of schools, but in the present case there were distinct variations in the nature of the hypotheses (an example is the specific hypothesis of teacher dominance in questioning, especially in the context of vocabulary teaching—see below). In addition we made an explicit hypothesis that was implicit in the Mangere study which is identified as the first hypothesis (see below). The summary data from the systematic coding are shown in Table 6, but as noted above the hypotheses were developed from these data and qualitative analyses of the transcripts.

Table 6 **Mean exchanges per lesson (standard deviations) at the beginning of Year 1 (n = 15 teachers)**

Types of exchanges	Early Year 1 means and (standard deviations)	
Total exchanges	40.87	(22.53)
Text related	33.07	(19.97)
Vocabulary question T	14.93	(4.50)
Vocabulary comment T	9.67	(13.48)
Extended talk T	11.13	(9.21)
Extended talk C	5.87	(5.33)
Text check T	9.07	(7.31)
Text check C	6.93	(6.26)
Incorporation	6.40	(3.62)
Awareness strategy	15.31	(8.26)
Awareness other	9.00	(6.50)
Feedback—High	24.53	(19.14)

Fine-tuning existing approaches

The first hypothesis was that existing activities and approaches could provide a vehicle for effective teaching. The general organisation of daily sessions, with a balance between large-group and small-group instruction with specific approaches provided a workable basis for more effective instruction. The observational data showed generalised engagement and generally suitable patterns of interactions within the format of the approaches. Perhaps consistent with this, the achievement data which was relatively stable across year levels (see above) suggested teaching was sufficient to maintain progress at close to expected rates over a year.

However this did not provide the acceleration needed. The observations suggested a need to clarify goals and purposes and adjust the approaches to optimise these goals. For example, if the functions of reading to students are to build vocabulary and increase acquisition of more complex language forms (see below), as well as exposure to particular topics and concepts, then the structure (format) and interactions can be fine-tuned accordingly. This would need to be done for each of the approaches including Shared Reading, Guided Reading (and other forms such as Reciprocal Teaching), and SSR. One important qualification to this general picture was that the classroom observations showed some considerable variation between individual classrooms. For example, with older level classes or special timetabling arrangements, a lower frequency of class reading sessions occurred. Four of the classrooms varied in frequency of shared reading (conducted weekly), and in two classrooms the guided and shared reading session occurred once a week.

That existing activities and approaches could provide a vehicle for effective teaching is indirectly indicated by the relatively high rate of teacher–student interactions during lessons (see Table 6). The analyses of classroom observations showed a high rate of exchanges on average over the classrooms, approximately one exchange every minute (41 exchanges in 46 minutes), although there was a wide difference across classrooms from one exchange every two minutes to almost 1.5 exchanges every minute.

A distinction was made between those exchanges that were directly related to text reading (for example they occurred with some reference to the text either as the text was being read or at the beginning or end) compared with exchanges that were not immediately related to the text (such as rehearsing strategies before any reference to the text to be read). This comparison provided an indication of the proportion of actual text reading. Close to 80 percent of the exchanges occurred in relationship to text reading, suggesting a high focus on texts.

Vocabulary instruction

Across all classes teachers were observed to identify and elaborate potentially new or unfamiliar vocabulary in both informational and narrative texts. The observations show teachers were elaborating words both outside of texts and in the course of reading texts, but the rate was variable and overall generally lower than optimal rates given the scores on the vocabulary subtest. Few instances of repeated practice of new words across reading, writing, and oral language were

observed, and there were few instances where new vocabulary was identified by children. Word charts or displayed records of new vocabulary, generated either by the teacher or children, were present in about half the classes. These were not observed being recorded or arising directly from a session, although in a Samoan bilingual class the teacher had vocabulary from the specific text prepared on cards. The good examples where vocabulary was elaborated included explaining idiomatic language and requesting text evidence for meanings. One example was an extended sequence where a teacher and children discussed what "stretch the money out" meant. Another example centred around the phrase "they had eyes but they couldn't see"—discussing how to interpret this seemingly contradictory sentence that was used figuratively. In another class teacher and students developed connotations for a word that is used in different curriculum contexts, and in another class teacher and students used knowledge of a word in Samoan to construct understanding of an English word.

An additional analysis of vocabulary instruction was undertaken in the present study. The transcripts were re-examined for every instance in which the teacher explicitly elaborated or defined a word. The elaboration or definition could include a synonym and might be as short as one word, but was judged to be explicit instruction if the teacher's contingent response did not simply repeat the word, but rather added new meaning about the word. These words were then categorised into three types. These were technical language words (e.g. 'prediction' or 'noun'); topic-related words, words directly related to the topic or theme that a text introduced or were part of a topic/theme (e.g. 'volcano'; 'comet'); and other words, which tended to be lower frequency words or idiomatic and figuratively used words and phrases (e.g., "disturbed", "peace of mind").

Many of the recorded instances of vocabulary interactions involved referring to and defining words, often in the form of questions (the data in Table 6 indicate that in 46 minutes of instruction about 15 of the total 33 text exchanges were exchanges which had vocabulary questions). The additional analysis of explicit teaching of types of words showed that averaged across the teachers 1 word per teacher was a technical word, 1.3 words were topic related and 2.9 words per teacher were other words. Terms were used for linguistic categories at subword, word, sentence, paragraph, and text levels and the exchanges often were associated with the teaching of strategies (see below) and the use of technical terms such as "clarifying", "predicting", and "visualising".

A second issue, made very obvious in the observations of strategy teaching, was that meanings of words were seldom checked in ways that elaborated specific connotations in context. Many of the instances involved discussion of students' ideas about meanings, often with the teacher accepting student contributions without critical appraisal. Few of the instances involved explicit instruction and modelling of how to check meanings within texts or via a dictionary or a thesaurus (see below).

These issues were fed back to the lead teachers who discussed them with their teachers in the form of a specific hypothesis that there was a need to increase the rate of vocabulary acquisition, especially nontechnical language, and in ways that gave access to multiple meanings and connotations. Research evidence to support the need to boost vocabulary through teacher

guidance in elaborations and feedback (e.g., Biemiller, 2001) was identified during this process. Possible ways that were discussed included increased use of reading to small groups of carefully selected texts which provided variation in genre and topic, with planned rates of exposure to new vocabulary. In addition, language acquisition research which noted how increased extended talk was associated with new vocabulary and with greater understanding of complex utterances was introduced (e.g., Hart & Risley, 1995). Other strategies included increasing the actual time spent reading either in groups or individually, because the records indicated a large amount of teacher discussion outside of texts.

The specific hypothesis for the teachers built on these patterns relating to vocabulary learning—the elaboration of specific vocabulary and extended talking during which new words or new meanings of words might occur. Elaborating words both outside of texts and in the course of reading texts and extended talk did take place, but the patterns had two problematic features. One was the presence during elaboration exchanges of high rates of teacher questions, often in the form of IRE (Initiate, Respond, Evaluate) sequences (Cazden, 2001). Questions were present in over a third of the total exchanges; they were almost double those in which the teacher commented on the meanings of words, and considerably higher than exchanges which looked at meanings in dictionaries (mean = 1.40 exchanges). The rate of questions was almost three times as high as the rate for teachers in Mangere at the beginning of their second year in the intervention (vocabulary questions teacher mean = 6.53 exchanges). A high rate of teacher questioning can reduce the complexity of children's learning, by dampening child contributions (Wood, 1998). In addition to this, few instances of repeated practice of new words across reading, writing, and oral language were observed. A second feature was the low number of exchanges with extended talk by the teacher. Generally, it has been found that the more extended talk by an adult or teacher, the more extended talk there is by the child, and this provides an important platform for learning aspects of complex language including vocabulary (Dickinson & Tabors, 2001). About 25 percent of the exchanges involved some extended talk by the teacher, and less than 15 percent of the exchanges involved students' extended talk, in both cases representing about one such exchange every four or more minutes.

This concern for vocabulary teaching was a deliberate focus in the Mangere study and the evidence showed that when teachers increased their vocabulary instruction, STAR subtest 4 (vocabulary) scores also increased. The specific hypothesis fed back to the teachers was that more specific instructional focus was needed which involved less dominance by teachers, more repeated practice of new words with increasing the particular focus on "other words". Given the large SDs, an additional hypothesis was made that more consistency between teachers was needed.

Strategies supported by checking and evaluating threats to meaning

In the Mangere study we developed a hypothesis about the instruction for reading strategies. In that study the deliberate reference to, teaching of, and use of comprehension strategies was

present in all classrooms (these were exchanges in which specific reference to building awareness of strategies occurred). But an issue with strategy instruction was suggested very early on in the observations in the first year. The issue was that in strategy instruction there was limited use of text evidence to detect confusions or threats to meaning, or to check and corroborate meanings. Additionally, this was observed as a limitation with vocabulary instruction (as noted above) and in the examples of incorporation (see below). There were few instances where the children were asked to provide evidence for their analyses, comments, or elaborations (such as "How did you know?"). This limited reference to texts to check understanding was especially noticeable with the use of predicting in the whole-class and small-group activities in which a text was shared, or introduced for some form of guided reading. In every such activity observed ideas were generated with high engagement. However, explicit direction to check the evidence to see if what was predicted was supported in upcoming text (at sentence, between sentence, and text levels) happened infrequently. Across all classes in all activities predictions were often prompted; for word meanings and for event outcomes and sequences, and without exception accepted and supported ("Good prediction", "That was clever", "Could be"). Most of the dialogue observed was about generating ideas, not checking them, and the teachers' responses were to accept and reinforce predicting. Similarly, asking for predictions often led to exchanges in which the students tried to figure out what the teacher was thinking about.

As in the Mangere study, the need for detecting threats to constructing accurate meanings and checking texts for evidence to support meaning is indicated by the achievement data, especially the low scores on the cloze test (subtest 3—paragraph comprehension). The observational data show areas of strength here but also generally lower-than-expected rates at which predictions and meanings of words, phrases, and passages are checked against text sources. Exchanges were analysed for the presence of detecting problems in meaning and checking evidence from texts. There needed to be some explicit reference to checking and evaluating evidence and this could be in the form of questions, directions, prompts, feedback, or comments as occurs in each of the examples above (e.g., "Anything else towards the end of the paragraph?"). Table 6 indicates a relatively low rate; teachers focused on checking in 9% of the exchanges, and only in 7% of the exchanges were children checking evidence.

The hypothesis fed back to the teachers was that comprehension would be enhanced with more direct explicit instruction and modelling of checking for evidence—for inferences, for meanings of words, for coherence etc—within sentences, between sentences, within a text, and even across texts. There is some research evidence that this could be a problem in strategy instruction leading to formulaic use of strategies in general and guessing rather than the appropriate use of texts to support inferences, to clarify meanings, to maintain coherence, and to predict (Baker, 2002). Similarly, recent research reports have identified groups of children who have fast efficient decoding but low comprehension, who have a high rate of errors termed "excessive elaborations", which are essentially guesses (Dewitz & Dewitz, 2003). This was a deliberate focus in the Mangere study and students made gains in scores on subtest 3 (paragraph comprehension).

Increased incorporation and awareness

Two complementary processes have been proposed as particularly important in effective teaching with culturally and linguistically diverse students (McNaughton, 2002). One of these is the use of students' expertise in classroom activities. At one level this involves capitalising on their event knowledge and interests through instruction including the selection and matching of texts. At other more complex levels this involves using familiar language forms and even types of culturally based forms of teaching and learning. But complementing this process is instruction that increases students' awareness of the relevance of their skills and knowledge and relationships with the goals and formats of classroom activities.

In each of these classrooms there were instances where teachers incorporated their students' event knowledge and language skills, drawing on their social and cultural identities. There are examples in the transcripts of teachers drawing on background knowledge. Careful selection of texts meant teachers were able draw on students' knowledge of Māori and Pasifika themes, images, and language; for example in two classes reading "The Plum Tree" by Paora Tibble (2002), and other classes reading "A Silent World" (Utai & Rose, 2002), or a newspaper article entitled "Outdoors Beckons Jonah Lomu". In one example, connections were made between elements of a Māori legend and local event knowledge, which generated considerable discussion and opportunities for language extension. This was an area of strength and perhaps is indicated in the achievement data by the variation in difficulty between passages on the PATs. But at a more complex level, there were few examples of teachers drawing on language and literacy skills embedded in familiar activities from outside of the classroom. Despite the rich examples, Table 6 suggests that the number of exchanges that contained incorporation of children's cultural linguistic or cognitive resources was relatively low; there were 6.4 exchanges in 46 minutes. The proportion of exchanges in which some incorporation occurred was 16 percent.

The complementary dimension, building the learner's awareness in classroom activities, was more problematic. Transcripts revealed quite wide use of learning intentions and their success criteria being made explicit. Similarly, the critical discussion and analysis of assessment formats was a strength. Specific use of some form of Reciprocal Teaching (Brown, 1997) was directly observed in 10 classrooms, and all schools reported some use of specific strategy instruction. Explicit instruction to be aware of the use of strategies often occurred; in terms of exchanges averaged in Table 6 there were 15.31 exchanges in 46 minutes, representing almost half of the text-related exchanges. However, a general question partly derived from the transcript analyses was the extent to which children understood the text-related goals of what had been made explicit. That is, there was a question as to whether children knew why the strategies of Reciprocal Teaching were to be used and how these serve overriding goals of engaging with and constructing meanings from the text. In the feedback to teachers it was suggested that the rates of instruction focused on priming students' awareness of strategies may have been too high. However, it was also suggested that, more generally, exchanges focused on building students' awareness of other aspects of performing effectively (other than specific strategies—"awareness other") was relatively low, occurring 9 times on average over 46 minutes.

The feedback with teachers focused on the hypothesis that students' learning would be improved if instruction enhanced students' awareness of classroom goals and formats, and their knowledge and skills in relationship to these. Vehicles for this might include the information provided in contingent feedback as well the setting of clear and consistent learning intentions (Hattie, 1999). In this respect the rate of high-quality feedback was high (in Table 6 the mean rate was 24.53 exchanges in 46 minutes). The hypothesis suggested fine-tuning feedback around developing awareness of what was required, and more particularly what was required to become even more effective.

The development of reading / learning communities, including increased exposure to texts and planned variation across texts

Taken as a whole, the observations suggested that current levels of interest and engagement in texts could be developed further. A means for doing this would be to draw on the resources provided by peer groups, or ability/achievement groups as communities of text readers. This would involve developing groups in which reviewing, discussing, and recommending texts was common practice. While the overall observations supported this argument, the need was also indicated in the distribution of the achievement data. There was a noticeable proportion of students who scored stanine 5 and above and extension of these children was also a concern. The role of the teacher as both a guide and a model would be crucial in such groups (Dyson, 1999).

The general hypothesis was that deliberate building of such communities would increase engagement in reading and reading mileage, affecting the extent of practice within texts and planned variation in exposure across types of texts. The instructional need is to increase exposure to texts and text-based language, which increases, among other things, understanding and use of all sorts of language including idiomatic language, knowledge of genres, general knowledge, and motivation to read for a range of purposes. The observations suggested some limitations in accessible texts immediately available in classrooms. That is, limited numbers and types of texts, for example fiction texts and picture books for topic use or recreational reading uses or SSR. There was no obvious access to and use of electronic texts (such as Learning Media's e journals). The texts available and in some instances used with groups might therefore be providing limited challenges in terms of new language forms and vocabulary for middle- and high-ability groups. On the other hand, sometimes texts selected might be too difficult or have little relevance for lower-achievement groups. There were very good examples where challenging but high-interest texts were read to the class, providing many opportunities for encountering new words. These included reading Harry Potter books with obvious enjoyment.

The research evidence suggests that instruction for "minority" students may inadvertently reduce the engagement in cognitively complex tasks, and tasks that are critical to the long-term development of reading comprehension (e.g. McNaughton, 2002). The identified risks therefore are around limited practice. The hypothesis developed with the teachers was that in each of the areas of concern instructional "density" could be increased. Testing this hypothesis would require

attention to the general textual resources in classrooms including electronic and internet-based resources; the link between home and schools; and attention to discourse features which increased rather than reduced engagement in text reading.

Summary

The observation records and teacher reflections were used, together with the achievement data, to develop a set of possible directions for increasing instructional effectiveness. In essence these were emerging hypotheses about how instruction might be formatted to be more effective for reading comprehension. The hypotheses were about more effective instruction.

A major alternative hypothesis to those developed here (but not incompatible with them), is that the children had difficulties comprehending because of limits in their accuracy and fluency of decoding. Apart from small groups within classes who were having special instruction (such as Rainbow Reading) the teachers generally felt accuracy and fluency was not a problem. They could refer to running records and sources of evidence for this. The pattern of results in the PAT and STAR are generally consistent with this. The word recognition subtest of the STAR had the highest scores and some later passages on the PAT were done better than earlier passages.

Longitudinal cohort analyses

The following analyses track the achievement of a cohort of students from the beginning to the end of the project (i.e., from Time 1, Term 1, 2004 to Time 6, Term 4, 2006). Analyses were only conducted with the same students who sat all six tests to avoid any confounding effects from students with differential exposure to the programme. The three-year timeframe with the focus on Years 4–8 and the demographics of the schools, (contributing primaries (Years 0–6) and intermediate) meant that the only students available to be tracked were those who began with the project in Year 4 (Cohort 1). The achievement of other students not represented in this analysis occurs in the following sections.

Overall gains in achievement for longitudinal cohort

There was a statistically significant overall acceleration in achievement from Time 1 to Time 6 of 0.85 stanine. Based on estimations from stanine gain, this represents about three-quarters of a year to a year's progress in addition to expected national progress over the three-year period (see Table 7). By the end of the project, the average student scored in the "average band of achievement" (mean = 4.01) whilst in the beginning of the project, the average student scored in the "below average" band.

The cohort made statistically significant accelerations in achievement across the three years. The effect sizes for the age-adjusted scores and raw scores are higher than reported in an international

study by Borman (2005) on schooling improvement initiatives, which reports effects of between 0.1 and 0.3[3] for schooling improvement initiatives that have been running for a short time, although Borman reports on a small number of studies of school improvement which cumulatively over more than seven years achieved gains with effect sizes of around 0.5.

Table 7 **Stanine and raw score means for Cohort 1 at Time 1 (February 04) and Time 6 (November 06)**

		Stanine					Raw scores				
		Time 1	Time 6	t value		ES	Time 1	Time 6	t value		ES
Cohort 1	Mean	3.16	4.01	6.77	***	0.64	16.38	32.14	27.13	***	2.41
(Year 4, 2004)	SD	1.17	1.48				6.16	6.92			
	N	98	98				98	98			

Figure 10 and Figure 11 present the overall changes for Cohort 1 (n = 98) in terms of the stanine distributions. Figure 10 displays the percentage of students in each stanine at the beginning (Time 1) and the end of the project (Time 6), Figure 11 shows the percentage of students in each of the achievement bands, and Table 8 provides the mean percentages in each of these bands. There was a marked reduction in the percentage of students at the below-average stanines (from 59 percent to 28 percent in the Below Average band, a decrease of 31 percent) and an increase in the percentage of students in the Average band (from 33 percent to 63 percent, an increase of 30 percent). However, there was little decrease in the very low (stanine 1) stanine bands, and very little increase in the Above-Average and Outstanding bands. This means that students' achievement gains were made primarily from shifting students from the Below Average into the Average bands. Because of the small numbers, no further statistical analyses were carried out on the stanine distributions. These patterns are very similar to those for the overall cohorts in the Mangere cluster at Time 1 and Time 6, except for the Above Average and Outstanding bands at Time 6. In the Mangere cluster these totaled 10 percent of the cohort, whereas in this study they were 3 percent at Time 6

[3] Of note is that the effect sizes for raw scores are often more than double that of the stanine scores. The difference is because the stanine effect size shows the effect of the intervention when the scores have been grouped into bands (4–10 raw score points in each band) and age adjusted against national norms. This provides information on the size of the effect adjusted against nationally expected progress, in short, the effect size for *accelerations* in achievement. By contrast, the raw score effect sizes shows the effect of the intervention without adjustments against national norms.

Figure 10 **Stanine distribution at Time 1 (Term 1, 2004) and Time 6 (Term 4, 2006) against national norms**

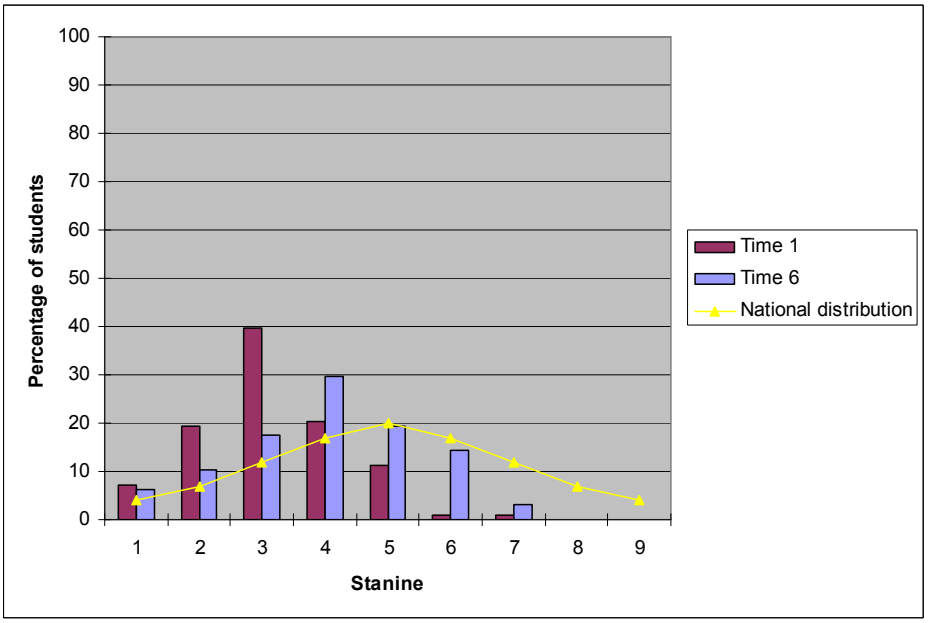


Figure 11 **Percentage of students scoring at Low, Below Average, Average, Above Average and Outstanding Bands at Time 1 (Term 1, 2004) and Time 6 (Term 4, 2006) against national norms**

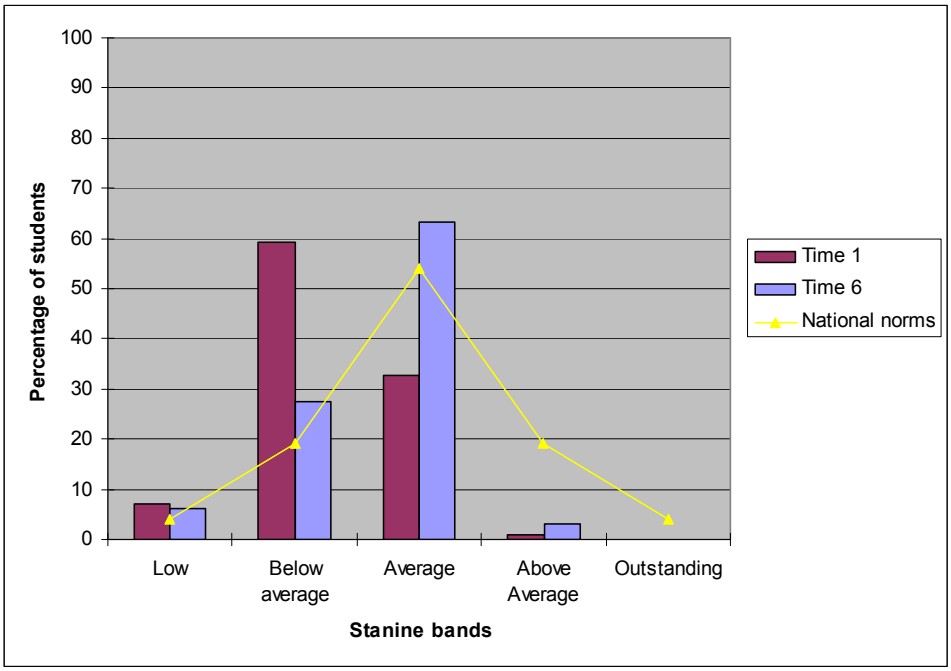


Table 8 **Mean percentages of students (and numbers of students) in stanine bands at Time 1 and Time 6 with expected percentages**

	Low (Stanine 1)	Below Average (Stanine 2–3)	Average (Stanine 4–6)	Above Average (Stanine 7–8)	Outstanding (Stanine 9)
Expected	4 (4)	19 (19)	54 (53)	19 (19)	4 (4) *
Time 1	7 (7)	59 (58)	33 (32)	1 (1)	0 (0)
Time 6	6 (6)	28 (27)	63 (62)	3 (3)	0 (0)

* The expected number of students in each band equals 99, rather than 98. This is due to rounding as the figures (to one decimal place) are 3.9, 18.6, 52.9, 18.6, and 3.9 respectively.

Gains in achievement across phases

As a whole, Cohort 1 made statistically significant accelerations (measured through stanine gains) in the first and second phase, but not the third (see Table 9). Gains were highest in the first phase and lowest in the last phase. However, each phase was associated with statistically significant raw score gains in achievement, but the gains were insufficient to produce significant increases in stanine levels in the third phase.

Table 9 **Stanine means and raw scores for Phases One, Two, and Three (Time 1–6)**

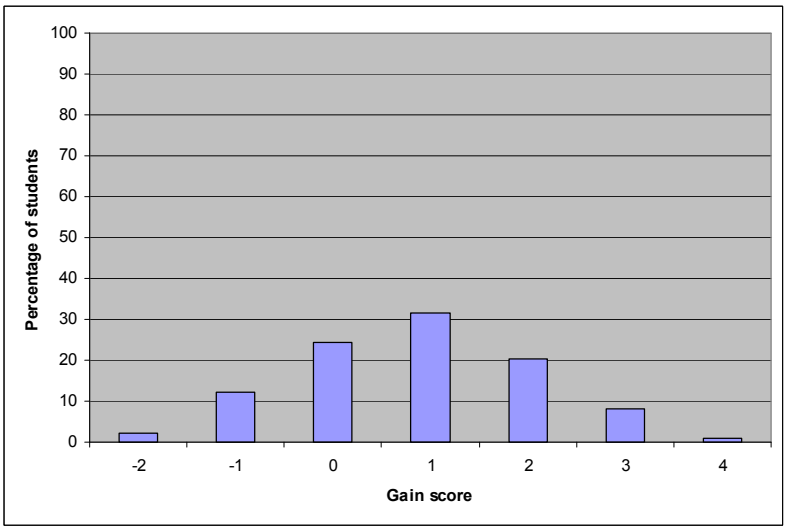
		Phase 1 (2004)				Phase 2 (2005)				Phase 3 (2006)			
		Time 1	Time 2	t value	ES	Time 3	Time 4	t value	ES	Time 5	Time 6	t value	ES
Cohort 1 (Year 4, 2004)	Stanine Mean	3.16	3.70	4.85***	0.42	3.66	4.00	3.50**	0.24	3.82	4.01	1.58	0.13
	SD	1.17	1.42			1.36	1.44			1.52	1.48		
	N	98	98			98	98			98	98		
Cohort 1 (Year 4, 2004)	Raw Score Mean	16.38	22.30	11.09***	0.89	22.77	27.73	10.27***	0.66	27.82	32.14	8.31***	0.59
	SD	6.16	7.10			7.55	7.43			7.60	6.92		
	N	98	98			98	98			98	98		

** p<.01
*** p<.001

Gain scores

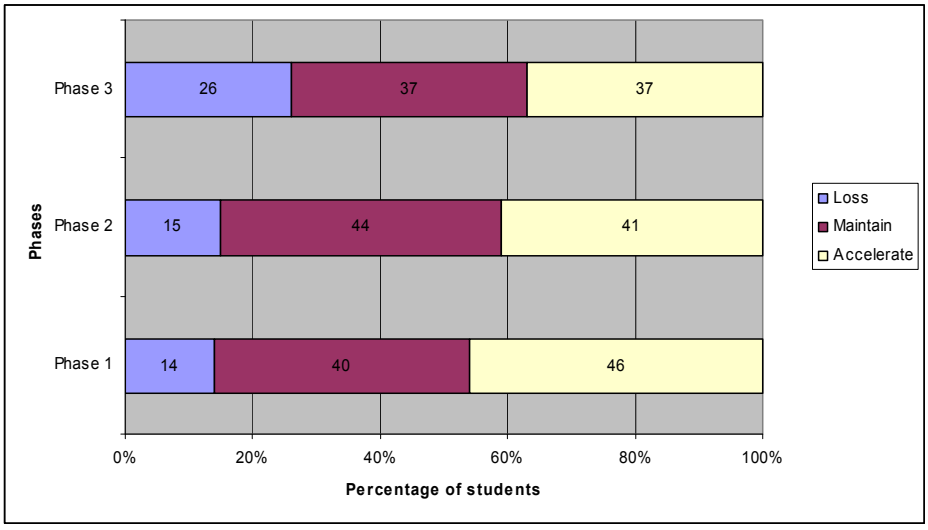
When compared to nationally expected progress, 86 percent of students in Cohort 1 maintained or accelerated their achievement from the beginning to the end of the project. The majority of students gained between one and four stanines (61 percent) or maintained their stanines from the beginning of the project (24 percent). Figure 12 shows the gain scores from the beginning to the end of the project.

Figure 12 **Gains scores from Time 1 to 6 for the longitudinal cohort of students**



The gain scores, when broken down by each phase, show that in each phase most students maintained or accelerated achievement over the three years (86 percent, 85 percent, and 74 percent respectively), with at least a third accelerating achievement in each phase (See Figure 13). Phase Three was associated with the most stanine losses and the least accelerations.

Figure 13 **Percentage of loss, maintenance, and acceleration across the three phases**



The achievement of Māori students

Māori students' achievement accelerated, like the other ethnic groups participating in the project (see Figure 14), gaining on average 0.91 stanine across the three years so that by the end of the project, the average Māori student scored within the Average band (mean = 4.29) (see Table 10). This means that Māori students made about a year's worth of progress in addition to expected progress over a three-year period. By contrast, at the beginning of the project, the average Māori student scored in the Below Average band. Other ethnic groups combined made slightly under a stanine gain (0.84 stanine) across the three years.

The intervention was designed from the profiles of the local students and their instruction, and contained elements that were designed to be both generic for the population of students and to be personalised using cultural and linguistic resources. It appears that the fine-tuning of instruction across the three phases of the research and development programme enabled this to happen.

Figure 14 **Mean achievement gain (in stanines) for Māori students compared to other ethnic groups combined**

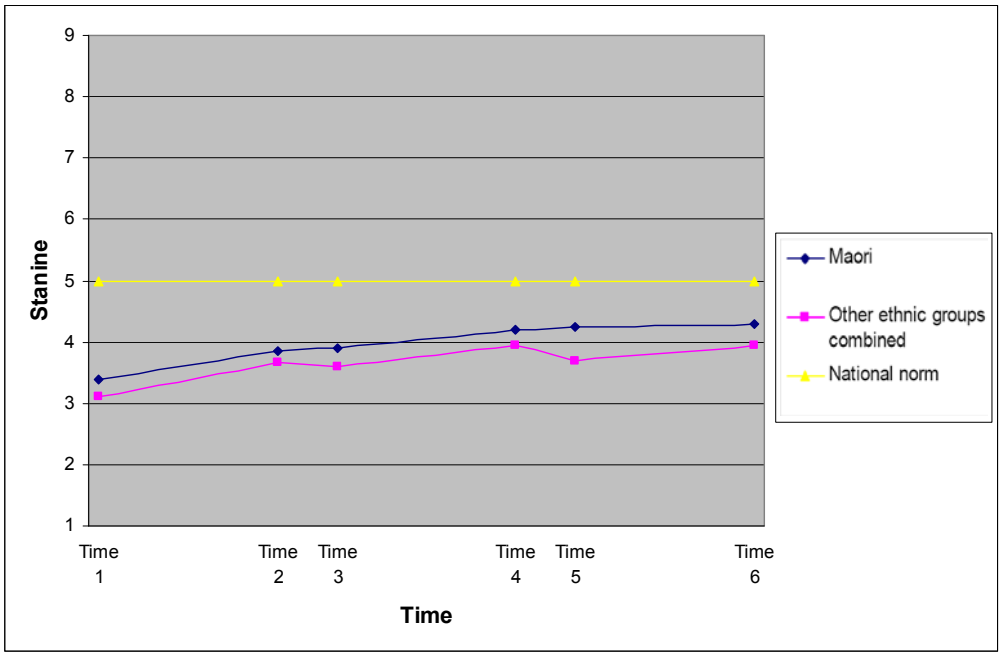


Table 10 **Stanine means by cohort for Māori students and other ethnic groups combined from the beginning to the end of the project**

			Phase 1 (2004)		Phase 2 (2005)		Phase 3 (2006)		Time 1–6	
			Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	t value	ES
Cohort 1	Māori	Mean	3.38	3.86	3.9	4.19	4.24	4.29	3.52**	0.69
(Year 4,		SD	1.16	1.46	1.51	1.4	1.51	1.45		
2004)										
		N	21	21	21	21	21	21		
	Other ethnic groups combined	Mean	3.1	3.66	3.6	3.95	3.7	3.94	5.79***	0.63
		SD	1.18	1.42	1.32	1.46	1.5	1.49		
		N	77	77	77	77	77	77		

* p<.05
** p<.01
*** p<.001

The achievement of males and females

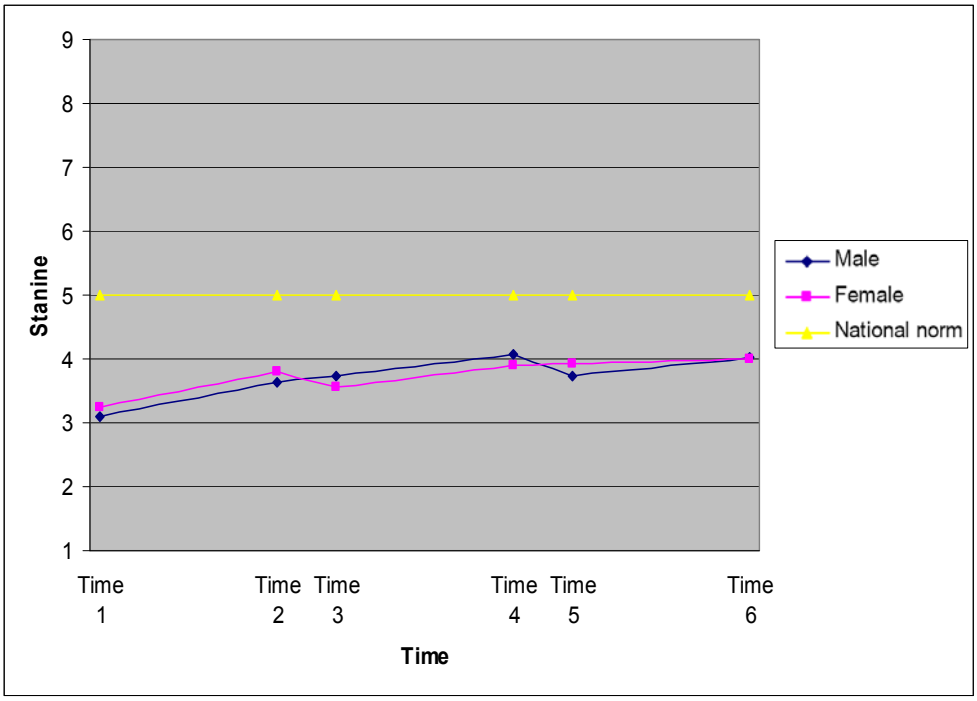
Overall, both males and females accelerated significantly, and by Time 6 both males and female students on average scored in the Average band of achievement compared with the beginning of the project when they scored in the Below Average band (see Table 11). Figure 15 shows that whilst both males and females started at different achievement levels, by Time 6 their average mean stanine was virtually the same.

Table 11 **Stanine means for Cohort 1 by gender—Phase One, Two, and Three (Time 1–6)**

			Phase One (2004)		Phase Two (2005)		Phase Three (2006)		Time 1–6		
			Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	t value		ES
Cohort 1	Male	Mean	3.10	3.64	3.74	4.07	3.74	4.02	5.71	***	0.61
Year 4, 2004		SD	1.33	1.56	1.48	1.55	1.58	1.66			
		N	58	58	58	58	58	58			
	Female	Mean	3.25	3.80	3.55	3.90	3.93	4.00	3.71	**	0.71
		SD	0.90	1.20	1.18	1.28	1.42	1.20			
		N	40	40	40	40	40	40			

** p<.01
*** p<.001

Figure 15 **Stanine means by gender—Phase One, Two, and Three (Time 1–6)**



School gains across the three phases

Only five schools had students we could track over three years. Each school was associated with statistically significant accelerations in achievement from the beginning to the end of the project, with effect sizes of between 0.47 to 0.88 (see Table 12). Three schools made gains of over a stanine, representing over a year's worth of progress in addition to expected national progress for a three-year period. The pattern across schools was of increasing acceleration in achievement from Time 1 to Time 6, although there was variation across schools in the amount of gain and in the pattern of losses and gains over the three phases (see Figure 16). We must, however, interpret these data in light of the small numbers in some schools.

Figure 16 **Stanine means by school—Phases One, Two, and Three (Time 1–6)**

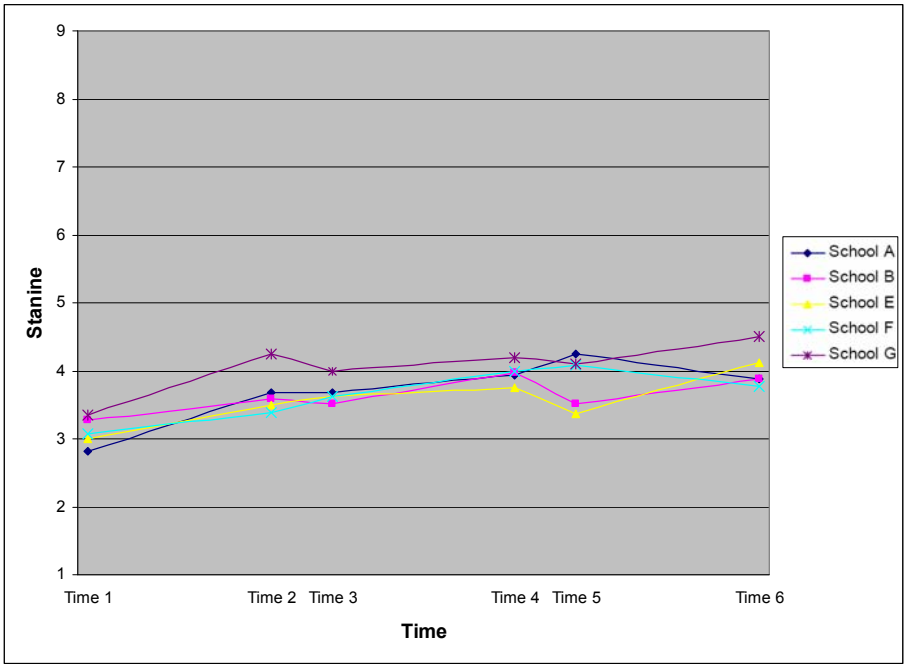


Table 12 **Stanine means by cohort for school—Phase One, Two, and Three (Time 1–6)**

			Phase One (2004)		Phase Two (2005)		Phase Three (2006)		t test Time 1–6		
			Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	t value		ES
Cohort 1	School A (N = 16)	Mean	2.81	3.69	3.69	3.94	4.25	3.88	3.44	**	0.85
(Year 4, 2004)		SD	1.22	1.45	1.35	1.39	1.57	1.31			
	School B (N = 41)	Mean	3.27	3.59	3.51	3.98	3.51	3.88	2.92	**	0.47
		SD	1.12	1.09	1.14	1.33	1.33	1.44			
	School E (N = 8)	Mean	3.00	3.50	3.63	3.75	3.38	4.13	3.81	**	0.88
		SD	1.07	1.69	1.60	1.75	1.60	1.46			
	School F (N = 13)	Mean	3.08	3.38	3.62	4.00	4.08	3.77	2.25	*	0.73
		SD	0.76	0.77	1.19	1.00	1.61	1.09			
	School G (N = 20)	Mean	3.35	4.25	4.00	4.20	4.10	4.50	4.20	***	0.67
		SD	1.50	2.07	1.81	1.88	1.71	1.91			

* p<.05
** p<.01
*** p<.001

Overall changes for total school populations year by year

A second way to analyse the achievement outcomes is to check the achievement of students within a year irrespective of presence or absence in any other year. This analysis answers the question "can changes in how the school teaches be detected in all children present in a year irrespective of being a member of a continuously present cohort?".

The following analyses examine improvements in achievement from the beginning (Term 1) to end (Term 4) of an academic year in each of the three phases. Analyses for each phase were only calculated from the same students who sat both tests in the academic year (beginning and end of year tests) to avoid any confounding effects from students with differential exposure to the programme. It should be noted that each phase built on the previous phase and included processes that were part of that phase. Common to each phase was the analysis, feedback, and discussion with teachers of evidence of teaching effectiveness. Note that there were differential participation rates between schools across the phases. One school could not participate in the first round of data collection, but participated in subsequent rounds of data collection. Another school participated in the Phase One and Two data collection, but did not participate in Phase Three as it pulled out of the project.

Overall gains in achievement

Table 13 includes all children present at both the beginning and end of each year of testing. The Term 1 and Term 4 comparisons show two things. The first is that each phase resulted in statistically significant gains from the beginning to the end of the year. This means that with the combination of continuing students as well as new students at each level, the effectiveness of the programme in accelerating student achievement in addition to expected national progress was sustained. Students made between approximately 15 and 20 months worth of progress for a year at school.[4] The improvements in mean stanine are shown graphically in Figure 17. The effect size in Phase One is higher than reported in international schooling improvement initiatives (effect sizes between 0.1 and 0.3 for under six years), and the effect size in Phases Two and Three are similar to those reported internationally (Borman, 2005). There were smaller numbers of students in Phase Three because one school only participated in the first two phases.

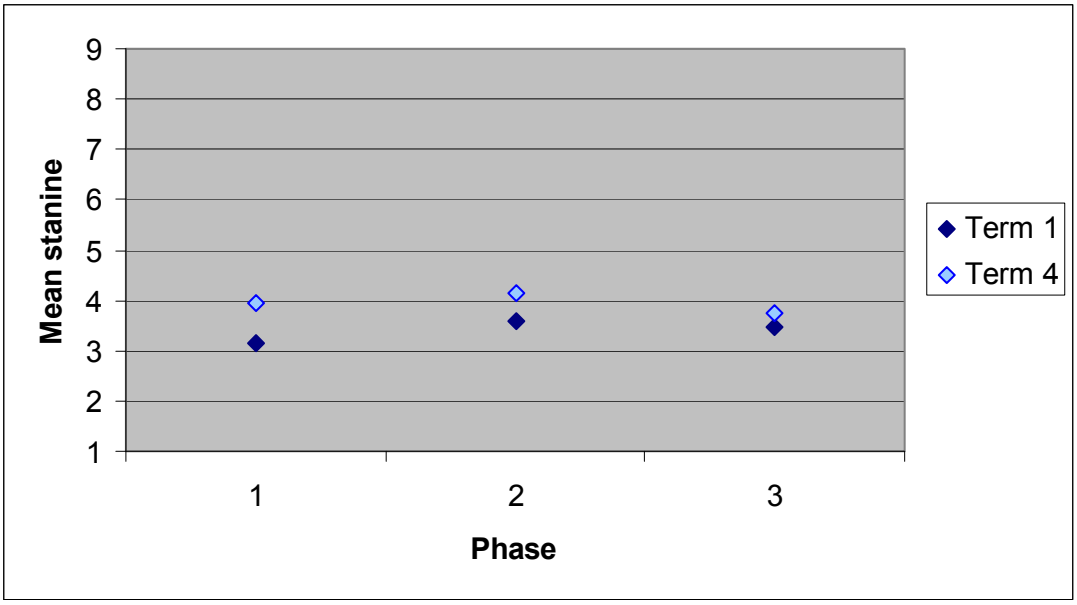
[4] Based on estimations from stanine gain.

Table 13 **Mean stanine and raw score in Term 1 and 4 in each phase**

	Term 1		Term 4			Gain	t value	Effect size
Phase One (n = 973)								
Stanine	3.14	(1.40)	3.95	(1.81)	***	0.81	17.21	0.50
Raw Score	26.08	(11.20)	36.13	(16.88)	***	10.05	29.33	0.70
Phase Two (n = 924)								
Stanine	3.59	(1.64)	4.13	(1.79)	***	0.54	12.19	0.31
Raw Score	26.44	(14.10)	32.98	(16.07)	***	6.54	24.57	0.43
Phase Three (n = 663)								
Stanine	3.46	(1.54)	3.74	(1.49)	***	0.28	6.58	0.18
Raw Score	21.91	(8.86)	26.61	(8.54)	***	4.70	24.47	0.54

* p<.05
** p<.01
*** p<.001

Figure 17 **Mean stanine for beginning (Term 1) to end (Term 4) of year in each phase**

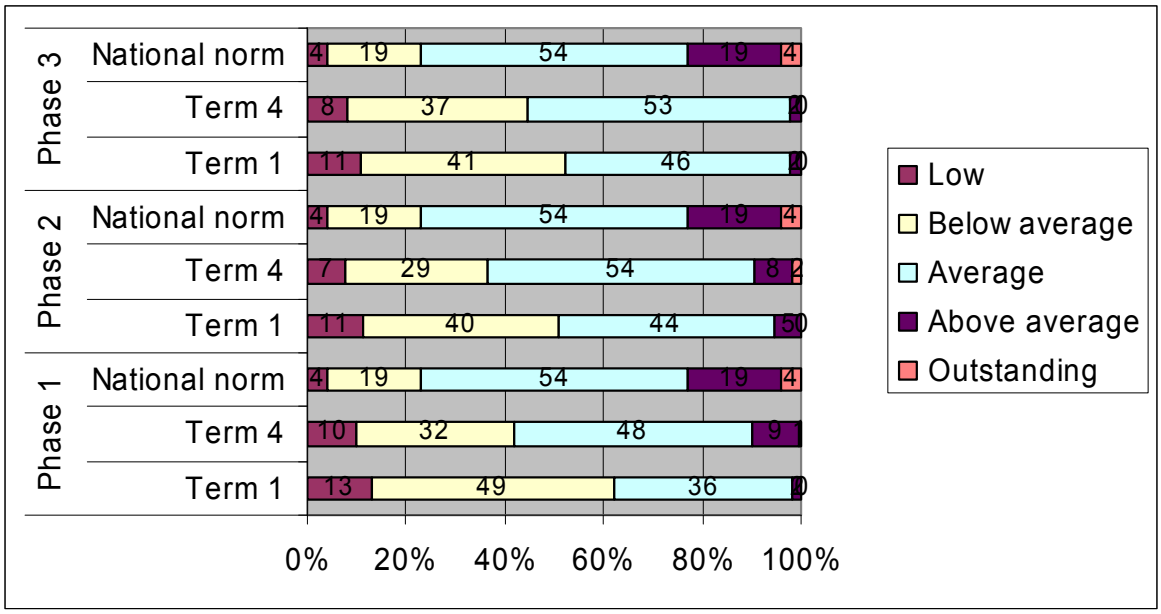


The second finding is that whilst the second year did not start at the initial level established in 2004 or finish at the levels in Year 1 (indicating that achievement levels tended to rise across those years), this was not the case for Phase Three. In Phase Three students began at a similar level to where they began in Phase Two and ended slightly lower than they ended in Phase 2. There are several possible reasons for this. For example, this could be because of the addition of transient students entering the school each consecutive year, which are characteristically different from students who were in the project from the beginning of the project. This and other hypotheses are elaborated at the end of this section.

Figure 18 shows the percentages of students in the achievement bands of Low (stanine 1), Below Average (stanines 2–3), Average (stanines 4–6), Above Average (stanines 7–8), and Outstanding (stanine 9). In each phase there has been a trend towards the nationally expected percentage of students in each band by Term 4.

Whilst there have been statistically significant improvements in mean scores (stanine and raw scores), increasing trends towards the nationally expected distribution and comparable effect sizes to successful international interventions, there are still fewer students in the Average and Above Average bands than nationally expected. As such, schools in the intervention still need to focus on improving student achievement further through sustaining their improved teaching and inquiry practices and developing new interventions to cater for the students who are now at and above average in their schools.

Figure 18 **Percentage of students in stanine bands in each phase (Term 1 to Term 4) compared to national expectations**



Year-level gains in achievement

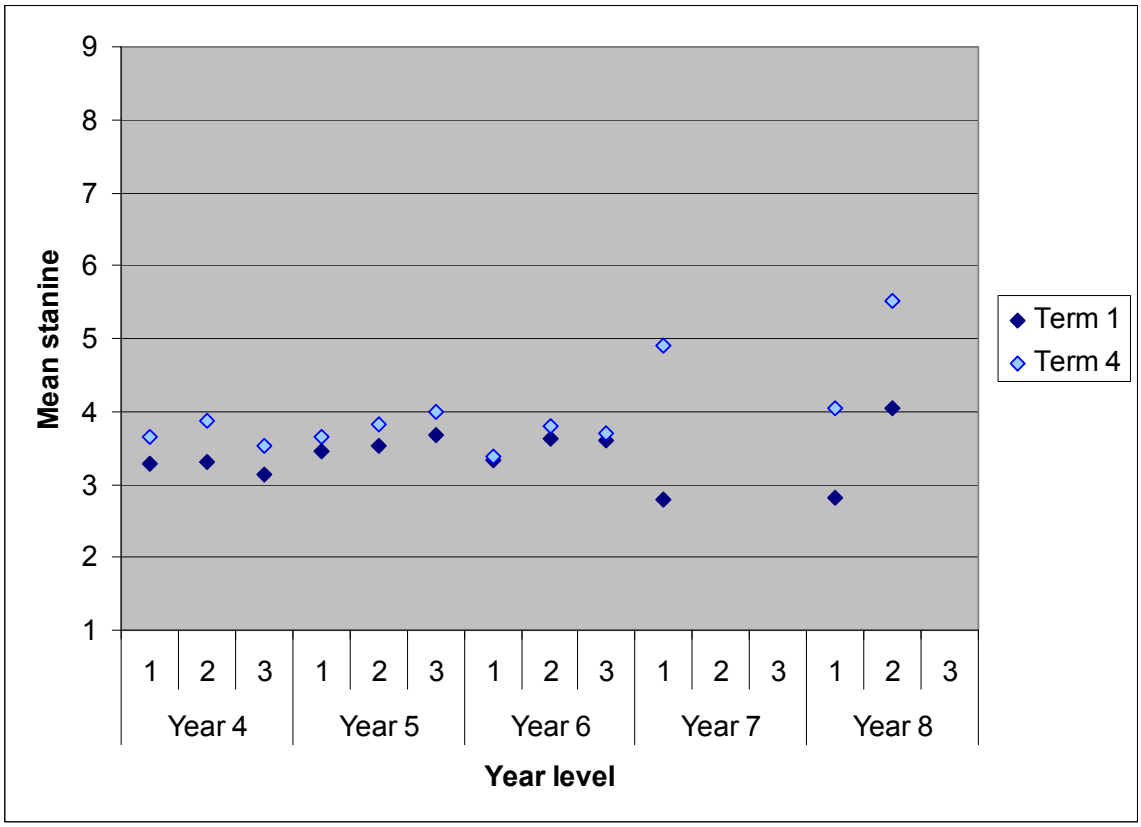
All year levels bar Year 6 made statistically significant accelerations in achievement in addition to expected national progress across all phases (see Table 14). (Note the school with year levels 7 and 8 pulled out of the intervention after Phase Two, and only gave Year 8 data in Phase Two.) However, in every phase, all year levels made statistically significant gains in achievement (as measured by raw scores) from the beginning to end of the year. In other words, whilst on average students in those year levels gained in raw score points, this was insufficient (when adjusted to nationally expected progress for a year) to significantly increase stanine scores for Year 6. Phase Two was associated with the least variability in achievement between year levels. The trends across year levels in each phase are shown graphically in Figure 19. It is interesting to detect the low gains in the Year 6 group. This parallels the longitudinal analysis which showed that the Year 4 cohort (Cohort 1) made low gains in the third phase when they were in Year 6. This suggests the pattern across years may be due to the year level rather than the phase per se.

Table 14 **Stanine means across year levels for Phases One, Two, and Three**

		Phase One					Phase Two					Phase Three				
		Term 1	**Term 4**	**t value**		**Effect size**	**Term 1**	**Term 4**	**t value**		**Effect size**	**Term 1**	**Term 4**	**t value**		**Effect size**
Year 4	Stanine	3.28	3.66	4.53	***	0.26	3.32	3.87	7.85	***	0.38	3.12	3.53	5.97	***	0.31
		(1.39)	(1.48)				(1.52)	(1.38)				(1.36)	(1.31)			
	Raw score	17.04	21.94	11.64	***	0.65	17.18	23.09	18.31	***	0.78	16.31	21.35	15.49	***	0.73
		(7.29)	(7.71)				(7.88)	(7.34)				(6.90)	(6.95)			
	N	174	174				256	256				226	226			
Year 5	Stanine	3.46	3.66	2.98	**	0.13	3.53	3.81	4.72	***	0.18	3.67	3.99	4.34	***	0.20
		(1.51)	(1.60)				(1.51)	(1.53)				(1.60)	(1.53)			
	Raw score	21.92	25.94	11.53	***	0.48	22.07	26.73	15.80	***	0.57	23.19	27.87	13.60	***	0.58
		(8.08)	(8.75)				(8.18)	(8.05)				(8.20)	(8.00)			
	N	217	217				255	255				229	229			
Year 6	Stanine	3.34	3.39	0.73		0.03	3.64	3.79	2.39	*	0.09	3.61	3.69	1.15		0.05
		(1.51)	(1.68)				(1.67)	(1.74)				(1.60)	(1.58)			
	Raw score	24.67	27.67	12.41	***	0.11	26.92	31.32	14.78	***	0.51	26.60	30.93	13.36	***	0.54
		(25.20)	(29.39)				(8.88)	(8.43)				(8.18)	(7.73)			
	N	193	193				244	244				208	208			
Year 7																
	Stanine	2.78	4.19	20.35	***	0.84										
		(1.22)	(2.02)													
	Raw score	30.19	51.28	26.35	***	1.59										
		(10.56)	(15.47)													
	N	216	216													
Year 8																
	Stanine	2.82	4.04	11.09	***	0.81	4.04	5.52	9.56	***	0.75					
		(1.21)	(1.76)				(1.86)	(2.10)								
	Raw score	36.24	51.01	21.64	***	1.26	46.36	59.76	12.51	***	0.92					
		(11.25)	(12.26)				(14.93)	(14.22)								
	N	173	173				169	169								

* p <.05
** p <.01
*** p <.001

Figure 19 **Mean stanine (Term 1 and 4) in each phase (1, 2, 3) by year level**



School gains across the three phases

All schools made statistically significant accelerations in achievement in Phase Two, with all but three schools making significant accelerations in achievement in Phases One and Three (see Table 15). One school did not accelerate achievement in either Phase One or Three. However, all schools made statistically significant gains in every phase as measured by raw scores, but this was not always sufficient (when adjusted to nationally expected progress for a year) to increase their stanine scores.

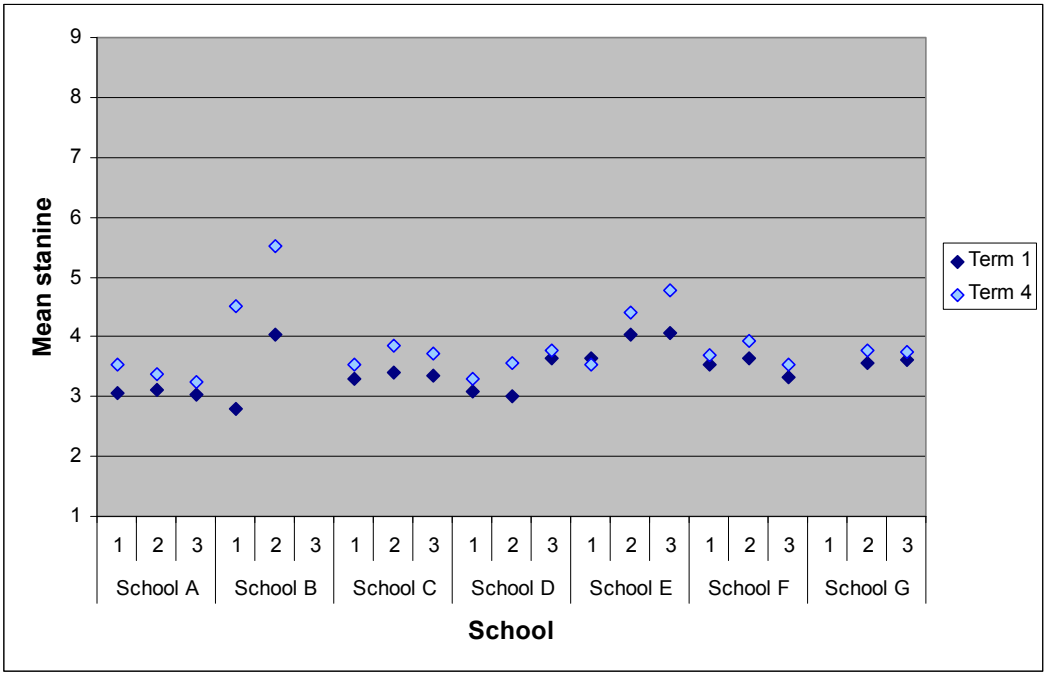
There was, however, a range of gains made between schools and within schools across the three phases. This suggests that schools may have differentially benefited from the combination of processes associated with the three phases. There did not, however, appear to be a Matthew effect operating where schools who were already succeeding gained more from the professional development (see Figure 20). School B made large gains although starting at the lowest level, and School E, with the highest beginning level in Year 1, made minimal gain in that year, and medium gains in the second year. This suggests that the impact of the professional development is mitigated somewhat by school characteristics.

Table 15 **Mean stanine and raw score (Term 1 and 4) for each phase by school**

		Phase 1					Phase 2					Phase 3				
		Term 1	Term 4	T value		Effect size	Term 1	Term 4	T value		Effect size	Term 1	Term 4	T value		Effect size
School A	Stanine	3.06 (1.51)	3.53 (1.79)	5.21	***	0.28	3.12 (1.29)	3.37 (1.41)	2.65	**	0.19	3.04 (1.59)	3.24 (1.38)	1.66		0.13
	Raw Score	20.22 (9.04)	25.68 (10.25)	12.75	***	0.56	19.90 (8.76)	24.70 (8.56)	10.50	***	0.55	19.85 (9.05)	24.16 (8.16)	8.41	***	0.50
	N	109	109				105	105				97	97			
School B	Stanine	2.80 (1.21)	4.52 (1.95)	21.77	***	1.06	4.04 (1.86)	5.52 (2.10)	9.56	***	0.75					
	Raw Score	32.88 (11.27)	51.51 (14.12)	33.01	***	1.46	46.36 (14.93)	59.76 (14.22)	12.51	***	0.92					
	N	389	389				169	169								
School C	Stanine	3.30 (1.34)	3.54 (1.48)	3.84	***	0.17	3.40 (1.48)	3.85 (1.52)	6.18	***	0.30	3.36 (1.59)	3.71 (1.49)	4.81	***	0.23
	Raw Score	21.16 (7.81)	25.84 (8.04)	15.63	***	0.59	21.37 (8.26)	26.99 (8.13)	17.38	***	0.69	21.18 (9.08)	26.48 (8.37)	15.52	***	0.61
	N	171	171				183	183				178	178			
School D	Stanine	3.09 (1.30)	3.30 (1.38)	1.57		0.16	3.02 (1.45)	3.55 (1.58)	3.84	***	0.35	3.63 (1.43)	3.78 (1.30)	1.16		0.11
	Raw Score	19.74 (8.40)	24.48 (8.33)	7.58	***	0.57	19.14 (9.39)	24.66 (9.24)	7.87	***	0.59	23.30 (7.58)	27.09 (7.03)	6.28	***	0.52
	N	46	46				56	56				54	54			
School E	Stanine	3.65 (1.50)	3.54 (1.46)	1.13		0.07	4.05 (1.59)	4.42 (1.48)	3.12	**	0.24	4.08 (1.28)	4.79 (1.28)	4.30	***	0.55
	Raw Score	23.04 (8.44)	25.79 (8.52)	5.27	***	0.32	25.38 (9.06)	30.15 (8.45)	8.81	***	0.54	25.42 (7.25)	32.30 (6.59)	8.98	***	0.99
	N	89	89				86	86				66	66			
School F	Stanine	3.54 (1.56)	3.70 (1.70)	1.66		0.10	3.65 (1.70)	3.92 (1.61)	3.05	**	0.16	3.32 (1.63)	3.54 (1.58)	2.09	*	0.14
	Raw Score	22.50 (8.90)	26.51 (9.24)	7.94	***	0.44	23.33 (9.65)	28.00 (8.49)	10.67	***	0.51	20.56 (9.89)	24.68 (9.80)	7.89	***	0.42
	N	169	169				153	153				102	102			
School G	Stanine						3.56 (1.63)	3.78 (1.56)	2.77	**	0.14	3.61 (1.45)	3.75 (1.43)	1.74		0.10
	Raw Score						21.91 (9.43)	26.62 (8.61)	12.72	***	0.52	22.90 (8.30)	26.93 (8.26)	12.00	***	0.49
	N						172	172				166	166			

* p <.05
** p <.01
*** p <.001

Figure 20 **Mean stanine in each phase by school**



Classroom gains across the three phases

The following figures (Figures 21, 22, and 23) show the gain scores in each classroom from Term 1 to 4 in Phases One, Two, and Three respectively. Phase One and Two had consistent gains across classrooms with 81 percent (38 out of 47) classrooms and 83 percent (39 out of 47) of classrooms maintaining or accelerating achievement respectively. In Phase Three 78 percent (28 out of 36 classrooms) maintained or made accelerations in achievement. (Phase Three also had the smallest number of classrooms that year because one school pulled out of the intervention.)

Figure 21 **Mean stanine gain score for classes in Phase One**

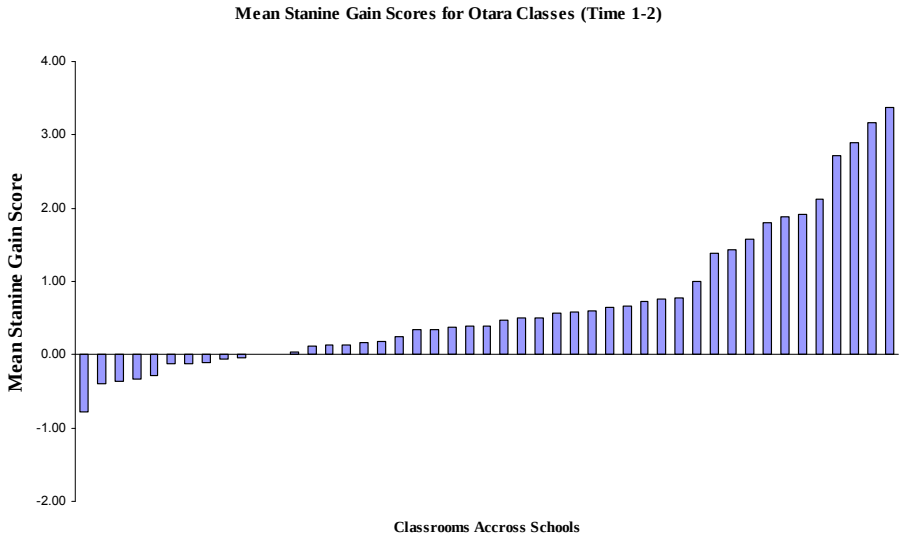


Figure 22 **Mean stanine gain score for classes in Phase Two**

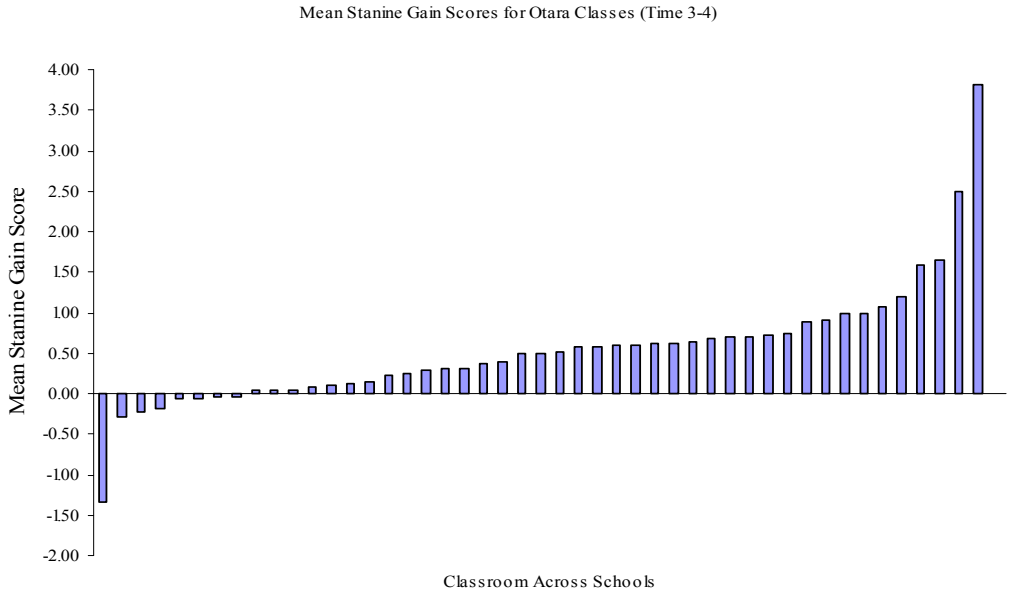
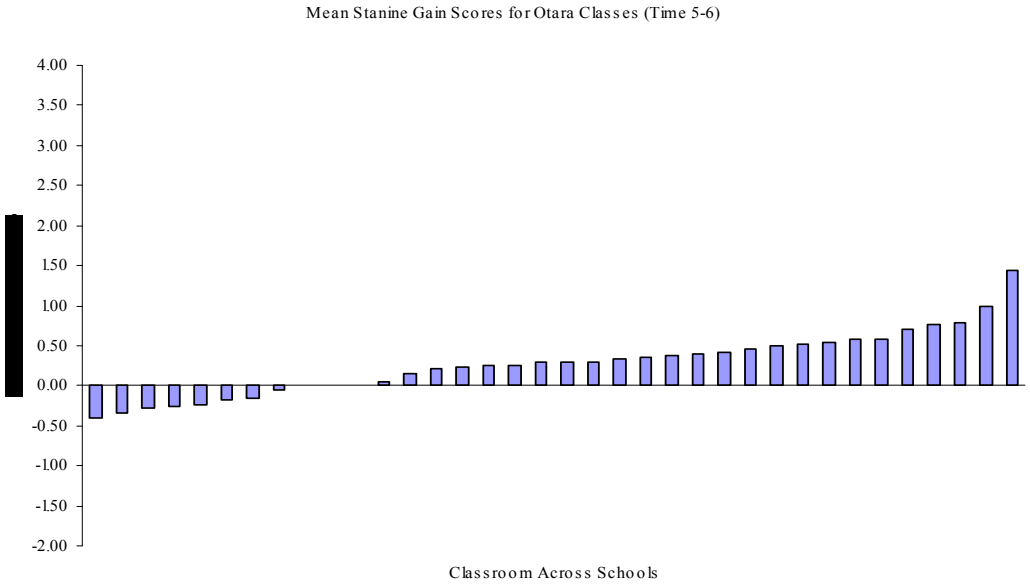


Figure 23 **Mean stanine gain score for classes in Phase Three**



As with the Mangere study we took some measures of participation and leadership in Phase Two and Phase Three as a preliminary approach to examining between-school variations. There are few consistent associations that this analysis reveals. The participation in the professional development sessions was rated as shown in Table 16. There was little difference between the schools in participation in the sessions. Some schools had a medium rating and some had high ratings but this difference did not appear to be related to the achievement gains. However the ratings of participation in the third phase, based on presentations at the teacher-led conference, do bear some relationship with the achievement gains. School E and School F participated strongly in the conference and both had significant gains in stanines in the third phase, although the only other school that had a significant gain in stanines did not make any presentations at the conference. Taking the two sets of analyses together one school, School E had consistently high ratings across Phase Two and Phase Three, and the achievement data at this school were consistently high, and even higher in Phase Three than in Phase Two, unlike any other school.

There were changes in literacy leaders and principals in some phases. For example, Schools A, B, and F had changes in principal in Phase Two in the second half of the year and School D had a change in principal in Phase Three.

Table 16 **Ratings of participation of staff and school leader in ten professional development sessions (Phase Two) by school**

School	Teachers[1]	Leader[2]
A	2	3
B	2	2
C	2	2
D	2	2
E	3	3
F	2	3
G	3	3

[1] 1 = fewer than 5 sessions; 2 = 5–7 sessions; 3 = 8–10 sessions
[2] 0 = did not attend; 1 = fewer than 5 sessions; 2 = 5–7 sessions; 3 = 8–10 sessions

Table 17 **Participation of school in Presentation of Inquiry Projects (Phase Three) by school**

School	Presentation[1]
School A	2
School B	(pulled out)
School C	0
School D	0
School E	3
School F	3
School G	1

[1] 0 = no staff presentation; 1 = one presentation (representing less than 50 percent classes); 2 = more than 1 but not all classes represented; 3 = all staff at all levels contributing.

Additional analyses: overall gains, all students all schools, and transient/absent student achievement

The final analyses present information on all students irrespective of continuing presence either within a year or over a year (see Table 18 and Figure 24). Note that different combinations of schools participated in the collection of data at any time point; for example, one school participated in the baseline data collection but subsequently pulled out of the intervention. This tells us about the performance of *all* students at any given time point. This shows that achievement had an upward trend despite the inclusion of students with differential exposure to the programme. The achievement levels at Time 6 are lower than for those students who had been

through the whole programme (mean = 4.01) (see Table 7). This suggests that students who had stayed in the same schools through the whole programme benefited more than those who had differential exposure to the programme.

Table 18 **Mean raw scores and stanines for all students from Time 1–Time 6**

	Raw score means			Stanine means		
	N	M	SD	N	M	SD
Time 1 (February 04)	1374	26.51	11.67	1374	3.02	1.40
Time 2 (November 04)	1108	35.18	16.74	1108	3.88	1.79
Time 3 (February 05)	1353	28.37	14.48	1353	3.54	1.63
Time 4 (November 05)	1071	32.83	16.10	1071	4.11	1.81
Time 5 (February 06)	848	21.99	9.00	848	3.46	1.56
Time 6 (November 06)	814	26.12	8.75	814	3.67	1.52

Figure 24 **Mean achievement scores (stanine) of all students at all time points**

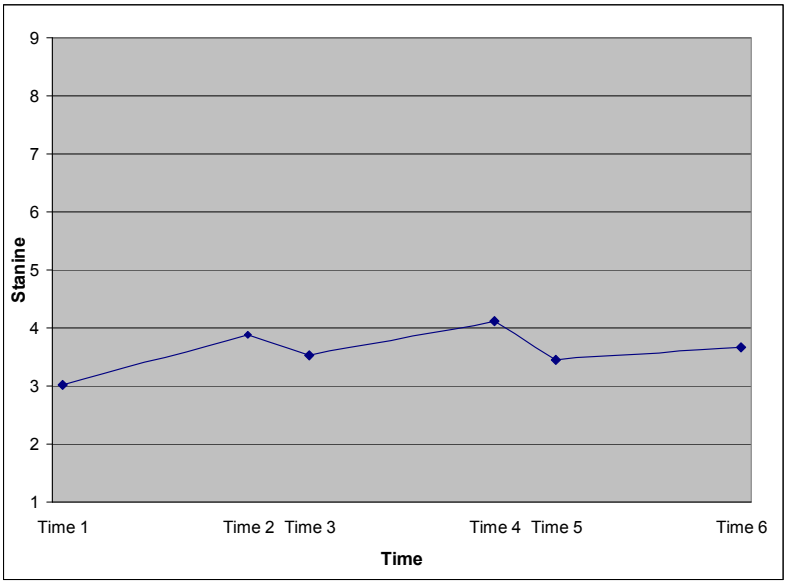


Figure 24 shows that for the data on all students, scores at the beginning of the year (Time 3, Term 1 2005; and Time 5, Term 1 2006) were lower than scores at the end of the year. As it is possible that the results are affected by new students coming into the schools at the beginning of the year, or students who were absent, we examined the results of students who were transient/absent compared to students who had completed 3 and 5 data points (i.e., students who fully participated in Phase One and Two respectively).

Our analysis examined the difference in scores at Time 3 (beginning of 2005) for students who were there from Times 1–3 (Phase One) and all other students at Time 3 (see Table 19). We then

repeated the analyses for Time 5 (beginning of 2006) for students who were there from Times 3–5 (Phase Two) (see Table 20) and Times 1–5 (Phase One and Two) (see Table 21) with students who were only there at Time 5. We excluded any Year 4 students who would be new to the project and therefore not absent or transient, and any Year 7 in an intermediate school who would be new to the school, and therefore not absent or transient.

The transience data suggest two conclusions. One is that the new children entering both clusters were similar to the baseline profile for each cluster, thus increasing confidence in the accuracy of the projected baseline. The second conclusion is that the new children were significantly lower than those who had gone through the intervention; up to one stanine in difference. This adds to the demonstration of the programme effect but also shows the constraints the schools were under to demonstrate improvements.

One hypothesis for the pattern of gains across phases might have been more transience in one phase versus another but there is no evidence in these data for that.

Table 19 **Comparison of absent/transient students against students who had completed all three tests at Time 3**

	Students with three tests			Absent/transient					
	N	M	SD	N	M	SD	t value		ES
Raw score	532	32.58	14.9	60	24.17	14.6	4.17	***	0.57
Stanine	532	3.8	1.7	60	2.8	1.65	4.34	***	0.6

* $p < .05$
** $p < .01$
*** $p < .001$

Table 20 **Comparison of absent/transient students against students who had completed the last three tests at Time 5**

	Students with three tests			Absent/transient					
	N	M	SD	N	M	SD	t value		ES
Raw score	432	25.33	8.14	49	21.45	9.86	3.09	**	0.43
Stanine	432	3.71	1.6	49	3.04	1.77	2.73	**	0.4

* $p < .05$
** $p < .01$
*** $p < .001$

Table 21 **Comparison of absent/transient students against students who had completed all five tests for current cluster (Otara) at Time 5**

	Students with five tests			Absent/transient					
	N	M	SD	N	M	SD	t value		ES
Raw score	121	27.83	7.32	49	21.45	9.86	4.64	***	0.73
Stanine	121	3.81	1.47	49	3.04	1.77	2.91	**	0.47

* $p < .05$
** $p < .01$
*** $p < .001$

Design-based longitudinal and cross-sectional comparisons

Like the Mangere study we have used a quasiexperimental design format to provide additional support for our judgements about treatment effects. In the Mangere study we used cohorts over one full calendar year (T_1–T_3) and two full calendar years (T_1–T_5) plotted against the baseline projections to make claims about effects and the replication across cohorts of those effects. In the present study we have adopted the same format and used the one-year data which provide us with three complete cohorts that we can plot against the baseline projections. In Figure 25 and 26, the cohorts in Otara and Mangere are plotted over one year. Figure 27 plots a cohort over two years, the only cohort in Otara that can be tracked over five data points. Visual inspection of the figures indicates that gains occurred in each cohort compared with the baseline projection. More formal statistical comparisons for the Otara cohorts are provided in Tables 22 and 23. The comparisons for the Otara cohorts are based on an adjusted baseline projection. The baseline has been modified to include only those students who were present over a year. This is to give a conservative projection given that we found that the students who were present only at the beginning of the year had lower mean scores. Tables 22 and 23 show that after one year, two cohorts (Cohort 2, Year 4–5 and Cohort 3, Year 7–8) had statistically significantly higher stanine scores than the baseline projections, but all cohorts had statistically significantly higher raw scores than the baseline projections.

Further analyses after two years show the Year 4 cohort was significantly different from the projected mean from the baseline for Year 6 (Year 4 Cohort stanine mean = 3.82, Year 6 baseline stanine mean = 3.34, $t = 2.77$, $p<.01$, ES = 0.32). This comparison is shown visually in Figure 27.

Figure 25 **Otara Time 1–4 cohorts against 2004 baseline**

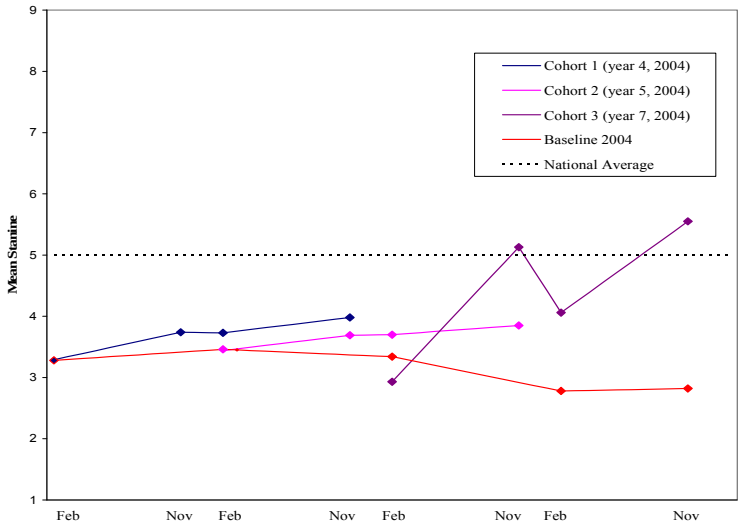


Figure 26 **Mangere Time 1–4 cohorts against 2003 baseline**

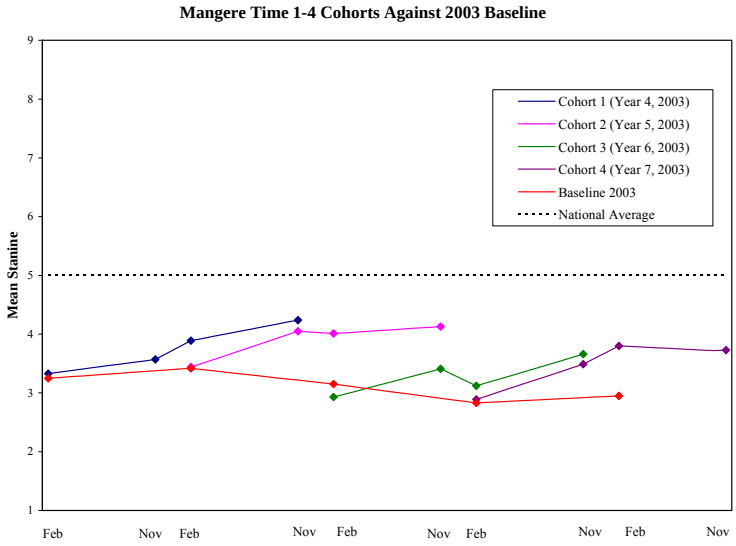


Figure 27 **Otara Time 1–5 cohorts against 2004 baseline**

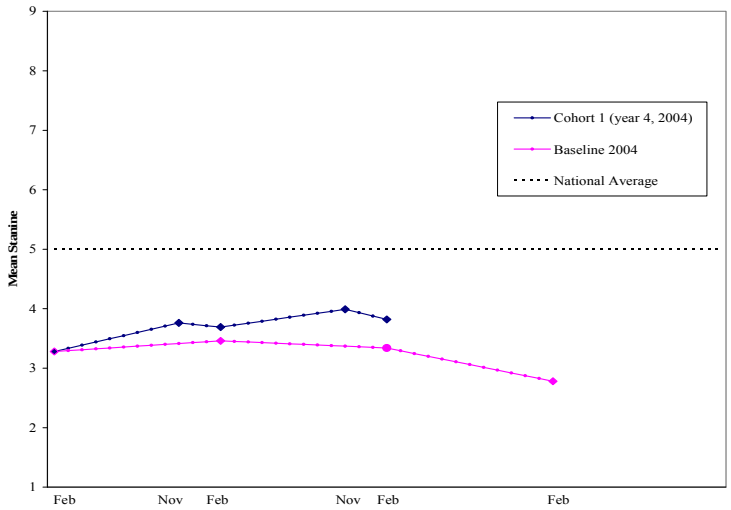


Table 22 **Stanine means by cohort for Otara baseline and Time 3 data**

		Otara Cross-sectional Baseline (Time 1, February 04)	Cohorts after one year of intervention (Time 3, February 05)	t value		ES
Cohort 1 (Year 4–5)	Mean	3.28	3.73	2.88	**	0.32
	SD	1.39	1.38			
	N	174	135			
Cohort 2 (Year 5–6)	Mean	3.46	3.7	1.49		0.15
	SD	1.51	1.64			
	N	217	165			
Cohort 3 (Year 7–8)	Mean	2.78	4.06	7.93	***	0.81
	SD	1.22	1.86			
	N	216	151			

* $p < .05$
** $p < .01$
*** $p < .001$

Table 23 **Raw score means by cohort for Otara baseline and Time 3 data**

		Otara Cross-sectional Baseline (Time 1, February 04)	Cohorts after one year of intervention (Time 3, February 05)	t value		ES
Cohort 1 (Year 4–5)	Mean	17.04	23.15	7.21	***	0.83
	SD	7.29	7.52			
	N	174	135			
Cohort 2 (Year 5–6)	Mean	21.92	27.38	6.41	***	0.66
	SD	8.08	8.46			
	N	217	165			
Cohort 3 (Year 7–8)	Mean	30.19	46.63	12.35	***	1.27
	SD	10.56	14.93			
	N	216	151			

* $p < .05$
** $p < .01$
*** $p < .001$

An additional comparison between Mangere and Otara is provided here which is consistent with the quasiexperimental design logic, but uses all students present at all time points. Figure 28 plots the gains by the averaged cohorts over Time 1 to Time 6 but shown in "real" time. The project in Otara commenced one year after the project in Mangere, so Figure 28 plots the data for Otara staggered after the Mangere data. Trend lines (best fit linear regression) have been fitted to the data points. The linear trend in Otara was very similar to that for Mangere with a similar estimate of fit and an almost identical slope ($y = 0.0219x + 3.3977$ and $y = 0.0217x + 3.0992$). Because only one cohort is represented in the Otara trend we did a second comparison with just the Mangere Year 4 cohort (Figure 28). The trend lines are slightly different (Mangere $y = 0.0325x + 3.4094$; Otara $y = 0.0217x + 3.0992$), with the rate of change for Mangere Year 4 slightly higher. Interestingly, a similar pattern of a drop and lower gains in Year 6 is indicated. (In each case the penultimate and final data points are Year 6).

Figure 28 **Mangere and Otara stanine means Time 1–Time 6 for students present at all time points**

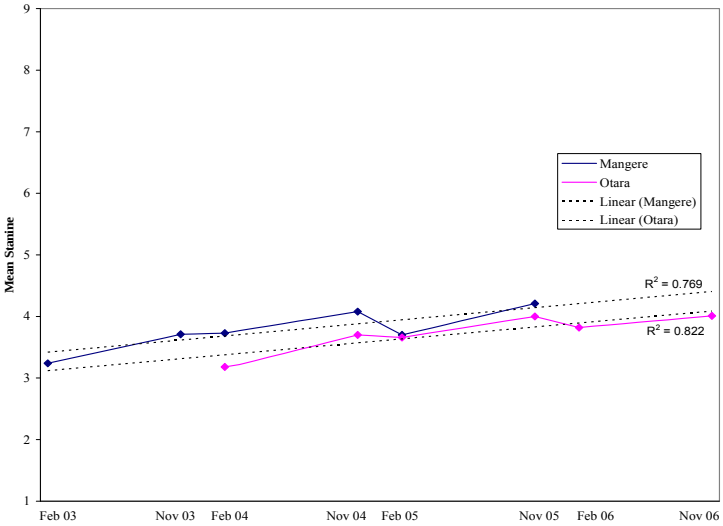
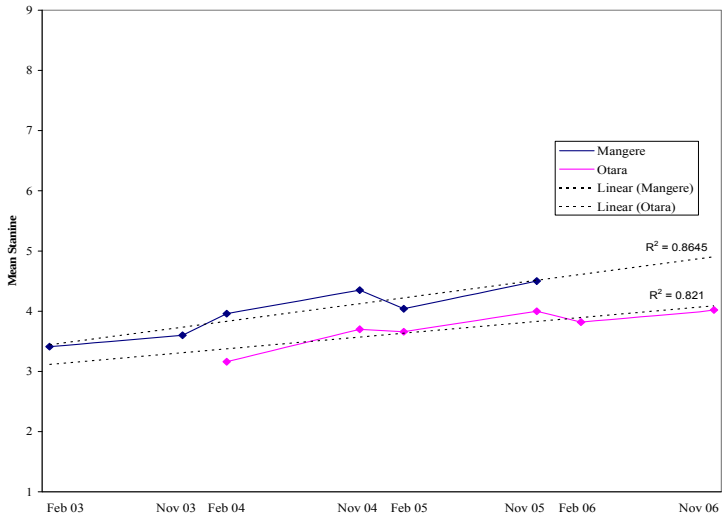


Figure 29 **Mangere and Otara stanine means Time 1–Time 6 (Year 4 only)**



Instructional observations (all teachers) 2003 and 2005

The observations taken in classrooms in the present study use qualitative and quantitative data to achieve two objectives. One is to establish treatment integrity or fidelity. Essentially this is the question: did the changes in instruction that take place actually match the focus of feedback and discussion in the first year and the development through professional development in the second year? The second objective is to plot the relationships between patterns of classroom teaching and the achievement patterns over the first two years.

Overall achievement gains and the general instructional focus over two years

As noted above the intervention resulted in statistically significant improvements across the first two years. These gains are summarised in Table 24 for the combined longitudinal cohorts who were at school for two years.

There was a significant increase in achievement between the beginning assessments (February 04) and the end (November 05) in every year cohort, with an overall gain of 1.24 of a stanine. The breakdown for the component tests is also shown in Table 24. The STAR test does not provide normalised equivalents for subtests so the raw score means are presented. Significant gains occurred across two years in all tests with very large effect sizes (using the raw scores), but a particularly large gain occurred in the paragraph comprehension (mean gain of 6.31 stanine). It should be noted that this test had 20 items, unlike the other tests which have 10 items, and the degree of gain may reflect the higher ceiling. The pattern is very similar for the combined cohorts in Mangere.

Table 24 **Mean gains in overall scores (stanines and standard deviations) in component tests (raw scores) across two years**

		Beginning (February 04)	End (November 05)	Gain	t value	ES
Total (stanines)		3.22	4.46	0.54	15.13***	0.74
	SD	1.34	1.95	1.14		
Tests (raw scores)						
	Decoding	7.25	9.45	2.2	22.19***	1.09
	SD	2.24	1.79	2.1		
	Sentence	4.47	7.32	2.85	22.88***	1.19
	SD	1.9	2.79	2.65		
	Paragraph	5.13	11.44	6.31	34.09***	1.57
	SD	3.55	4.45	3.93		
	Vocabulary	3.85	6.54	2.69	18.83***	0.98
	SD	2.07	3.28	3.03		

* $p < .05$
** $p < .01$
*** $p < .001$

Video records of seven classrooms were taken at the beginning of the second school year (February 2005). All seven teachers had been involved in the first year for the baseline profile, but only four of them had been observed in the first year. The video recording of classroom reading sessions was repeated at the end of the second year (November 2005). Seven teachers were again observed but only four of the original teachers who were observed at the beginning of the year

were available for observation a second time. In total that meant 10 teachers were observed at *either* the beginning or the end of the second year, four of whom were observed twice. The summary observations for their classrooms are shown in Table 25.

The seven teachers represented 18 percent of the total group of teachers in the second year (Phase Two). Their data can be used, albeit with some care, as indicative of teaching after the first year (Phase One) in the programme for two reasons. Firstly, they came from each of the schools and were in classrooms covering all the levels within schools. Secondly, the average gain made in their classrooms in the second year (Phase Two) (mean = 0.30 stanine), was similar to the overall gain for the other 40 teachers (mean = 0.37 stanine). However, their data needs to be treated with care because they are a mixture of teachers observed twice or once only at the beginning or the end.

The means per lesson for the exchange types observed in their classrooms at the beginning (February 05) and end (November 05) of the second year and early in the first year (February/March 04) are shown in Table 25. At the beginning of the second year the overall number of exchanges which occurred in 39 minutes were 37.29 exchanges, with a high density of these (32.86 exchanges, or 88.1 percent) being focused on a text, either during the reading of that text or in preparations or follow up which referred directly to the text. There was a high number at the end of the year (49.71 exchanges) and the observed time was 37 minutes.

The first comparison to note is between these means and the means for observations taken early in the first year. Two conclusions can be made, tentatively for the reasons noted earlier, and because these are small numbers of teachers. Hence we have not used planned comparisons. The first conclusion is that there is some evidence that changes in instruction consistent with the feedback to teachers and the associated hypotheses had occurred over the first year (see means for February 04 and 05). These included a consistent reduction in teacher dominance (indicated in the reduced frequency of exchanges with teacher questions and teacher comment); an increased focus on vocabulary through the vehicle of extended talk by both teachers and students; a reduced focus on deliberately instructing awareness of strategy use; an increased instructional focus on building students' awareness in areas other than strategies; and maintaining or increasing levels of high-quality feedback. Average levels of teacher and student checking in the second year (as seen in the means for November 05) were similar to the first year levels despite an initial increase, which suggests limited (in the sense of ongoing) change in this area.Levels of incorporation averaged at the start of the second year were similar to those in the first year, but there was an increase in the sample of observations at the end of the year.

A subsidiary analysis of vocabulary provides some confirmation that the focus on vocabulary was changed consistent with the feedback and instructional hypotheses. The explicit teaching of Other words (not Technical words or Topic Related words) increased to 4.4 words per teacher in reading sessions (averaged across the seven teachers) at the beginning of the year and was 4.1 words per teacher in reading sessions averaged for those seven teachers observed at the end of the year. Teaching of Technical and Topic Related words remained at around 1 word per teacher at both

times. Further analyses of the words reveals that teachers used a lot of the strategy-related words (e.g. "summarising"), but embedded in the instruction without explicit defining. The interactions in the transcripts below show the students knew what was required without needing to have the strategy explained.

Table 25 **Mean exchanges (and SD) early in Phase One, at the beginning of Phase Two and at the end of Year 2 (N = 7 teachers)**

	Early Phase One February/March 04		Beginning Phase Two February 05		End Phase Two November 05	
Text related	33.07	(19.97)	32.86	(12.75)	45.29	(12.16)
Vocabulary questions	14.93	(4.50)	8.29	(6.34)	6.29	(6.63)
Vocabulary comment T	9.67	(13.48)	8.71	(5.50)	4.14	(3.98)
Extended talk T	11.13	(9.21)	12.57	(6.27)	24.57	(8.81)
Extended talk C	5.87	(5.33)	12.57	(6.13)	8.86	(4.74)
Text check T	9.07	(7.31)	14.0	(11.56)	4.57	(4.43)
Text check C	6.93	(6.26)	10.86	(9.41)	2.86	(2.55)
Incorporation	6.40	(3.62)	4.86	(4.94)	10.71	(4.82)
Awareness strategy	15.31	(8.26)	8.43	(5.06)	3.86	(3.63)
Awareness other	9.00	(6.50)	28.43	(15.39)	14.00	(7.35)
Feedback high	24.53	(19.14)	25.29	(9.27)	28.71	(11.67

This pattern of change is consistent with the patterns of gains over two years in each of the component tests shown in Table 24. Of interest, however, is the gain on the decoding test which also increased by about the same degree (as shown by the effect size), and yet deliberately this was not a direct target of the intervention. Increases in decoding as a result of a well-targeted comprehension programme have been noted in the literature before and are likely due to the density effect of increased exposure to texts and the effects of more reading practise across more texts (Lefevre, Moore, & Wilkinson, 2003). The pattern of gains for students in the classrooms of the seven teachers was the same as the general pattern over two years. These were large increases over the second year in each component test (with effect sizes ranging from 0.87 to 1.10).

A further check on the pattern of change was made by examining the change over time for the four teachers who were observed at all three time points. Their data are shown in Table 24. The data in Table 26 indicate that the changes in exchanges for these teachers were consistent with the changes in the averages for the seven teachers in Table 25. Together the two sets of data (which contain the four common teachers) increase our confidence in concluding that instructional changes occurred which were consistent with the instructional hypotheses. The focus on vocabulary shifted from a question-and-answer format to more of an extended discussion, within which occurred an increased focus on new or unfamiliar words in addition to Topic Related or

Technical words. The evidence also suggests the teachers increased their focus on building students' awareness of tasks and their knowledge and performance other than specific strategies (they initially had high levels of explicit teaching of those strategies but reduced them over the second year). They increased levels of high-quality feedback even further. Their levels of incorporating students' background knowledge increased over the year.

Table 26 **Mean exchanges (and SD) for four teachers at three time points (beginning Phase 1 and beginning and end of Phase 2)**

All teachers (n=4)	Means (SD) Time 1	Means (SD) Time 2	Means (SD) Time 3
Text related	26.75	37.00	34.25
	(19.52)	(11.16)	(14.99)
Vocabulary questions	9.25	9.00	9.75
	(8.50)	(7.12)	(3.20)
Vocabulary comment T	6.75	8.25	5.25
	(8.89)	(7.27)	(2.63)
Extended talk T	10.75	15.75	21.25
	(6.29)	(3.77)	(11.18)
Extended talk C	6.75	15.75	17.25
	(5.73)	(3.77)	(10.31)
Text check T	6.00	18.00	12.50
	(2.58)	(13.09)	(7.33)
Text check C	4.25	14.25	10.25
	(5.31)	(10.56)	(6.60)
Incorporation	4.25	6.00	8.25
	(1.71)	(6.21)	(3.95)
Awareness strategy	13.75	11.00	1.00
	(7.23)	(4.24)	(.000)
Awareness other	2.75	32.75	20.75
	(1.50)	(18.30)	(6.02)
Feedback high	17.25	28.50	29.50
	(12.28)	(6.61)	(14.06)

One further analysis compared the overall pattern of teachers' instruction for Otara teachers in the second year with the patterns for the second year of instruction in the Mangere data. All 10 teachers for whom we had data on at least one time point in the second year for Otara were

compared with the nine teachers averaged over the second year in the Mangere study. In the case of Otara teachers who were only observed once the single measure is included, while averages were used for teachers observed twice. The data in Table 27 show quite similar patterns of exchanges. The only difference occurred in the category of exchanges which was focused on student awareness of strategies. Otherwise, the data suggest a replication of the instructional change occurred. That is, the intervention in Otara resulted in similar changes in instruction to those that occurred in Mangere.

Table 27 **Mean exchanges (and standard deviations) over Phase Two for Otara and Mangere teachers**

	Otara teachers (n = 10)		Mangere teachers (n = 9)	
Text related	29.75	(12.26)	25.89	(10.13)
Vocabulary questions	7.10	(4.49)	10.78	(7.52)
Vocabulary comment T	5.95	(3.65)	7.39	(6.20)
Extended talk T	14.70	(7.54)	11.89	(7.03)
Extended talk C	12.10	(7.01)	10.39	(5.95)
Text check T	11.80	(7.60)	8.00	(6.31)
Text check C	9.00	(6.62)	6.56	(5.49)
Incorporation	5.70	(3.81)	5.33	(4.14)
Awareness strategy	4.75	(5.25)	13.39	(9.84)
Awareness other	24.55	(11.03)	18.22	(11.53)
Feedback high	24.20	(10.09)	20.50	(8.95)

Case studies: high-gain and average-gain teachers

The achievement gains across the second year were used to identify two teachers who provided case studies of classrooms in which gains were larger than expected gains (that is gains expected for the pre and post time period). One was a very much larger than expected gain and the other was a large but average gain relative to the gain for the group of teachers as a whole across the second year. The mean gains for students in their classrooms on the component tests are shown in Table 28. Students in Teacher 1's classroom made gains in each component test, with particularly large gains in paragraph comprehension. Students in Teacher 2's classroom had slightly lower or similar gains in all areas except subtest 3 (paragraph comprehension), which was 2 raw score points less on average than students in Teacher 1.

Table 28 **Gains in Phase Two by two teachers on component subtests (STAR)**

	Stanine	Component Tests (raw scores)			
		Decoding	Sentence	Paragraph	Vocabulary
Teacher 1	0.9	0.41	0.78	4.46	0.44
Teacher 2	0.31	0.64	0.74	2.41	0.59

Case study 1: high-gain classroom

Teacher 1 taught in a composite Year 5/6 classroom. Children in this classroom made the fourth highest average gain in the second year (0.90 stanine). Her teaching was similar to the general approach with use of five small ability groups for guided reading (four days a week) and for shared reading in a fully structured session 2–3 times a week, but with additional reading to the whole class every day after every break. In the first part of the year the book read during the morning was *The Silent One* and in the afternoon after breaks she used sophisticated picture books. The whole class came together on Mondays on a new activity or concept. While she worked with particular groups other groups were engaged in reading tasks, including use of CD-ROM reading, peer reading, and learning centres.

This teacher's overall style was relaxed with clear indicators of her enjoyment of the children through exchanges with humour (discussing boat experiences she says: "Did they throw you off the side and you had to swim back?"; or when there are not enough books to read—"One of you lucky people is gonna have to share with me."). Her teaching included extended discussion in which students and teachers focused on meanings, with little explicit attention to teaching strategies, and careful use of texts drawing on background knowledge.

The use of strategies was apparent embedded in the text discussion, shown in the following example of inferencing during guided reading:

> T: Yes, digging a hole. OK. Does it have to be deep?
>
> C: [several] No.
>
> T: How do you know that? Point to, can you point to the part that tells you it doesn't have to be very deep? Don't look at her. You do your own one. Point to the part that tells you it doesn't have to be very deep.
>
> C: [one child says] Miss it doesn't it take long to dig.
>
> C: [another child reads] xxx it doesn't take long to dig.

The focus on vocabulary was readily apparent. There was explicit attention to new or unusual vocabulary, which was elaborated with shades of meaning and the meaning finally agreed. A lot of exposure occurred in the second sample when the teacher was reading *The Lion, the Witch and the Wardrobe* by C. S. Lewis.

The following example illustrates the process:

T: OK do you know what this word is? Michael's asking. "Counting's a hassle". What's a hassle? What does it mean?

C: Um it's hard.

T: It's hard.

C: [another child] A problem.

T: Or it's a problem. S?

C: Something you want to forget about.

T: Something you wanna forget about.

C: You might not remember.

T: You might not remember. What it means is counting's a… because they've got to count how many people, it just gets to be a pain after a while.

Other examples of explicit teaching of word meanings (other than Topic Related words and Technical words) included the words: "assembly line", "insert", "demonstrate", and "designed especially for kids". With the word "sturdy" examples of repeated practice occur within a guided-reading session. Four instances of use occur in the initial exchange, which was 11 turns long with the teacher signing off on the meaning. Following that, the teacher used the word in four separate exchanges, the first three reminding the class about its meaning and the fourth time just embedded in a new description. The students were engaged in these discussions and initiated some. In one instance the word to examine was "unfurl" but a student asked first "What does 'furl' mean?" To which the teacher responded. "OK so maybe we should find out what 'furl' means.…"

The teacher conveyed expectations through feedback which was both corrective feedback ("You're not answering my question") as well as being explicit about higher order thinking.

T: Cos you're thinking aren't you? OK, I should see you flicking back through the article making sure that you are correct, that you are self-checking what you are saying in your answer, OK? Now you're all probably gonna have different answers so don't worry if someone says something that's totally different to yours. OK, F, what do you think?

C: Cos if they started straight in the water they won't know what to do.

T: OK, so they wouldn't know what to do in the water.

In addition there was evidence of high levels of engagement by the students and a working community of readers. The markers of this were apparent in the reaction to having to finish reading *The Lion, the Witch and the Wardrobe.* With concern one child queried:

C: Finished the chapter?

T: Not quite. OK, we'll find out what happens later. Don't sneak the book.

C: [another] He's still alive!

Case study 2: Average gain classroom

The second teacher, who was at a different school from the first teacher, was also a mainstream teacher in a Year 6 class. On average students had made the same progress as that made by the whole group of students over the year. Their gains of component tests, shown in Table 28, were largest in paragraph comprehension (cloze test). She too followed the general approach with 3–4 small ability groups for guided reading (four days a week) and for shared reading 3–4 times a week, but with additional reading to the whole class every day, sometimes two or three times a day for 5–10 minutes. While working with a particular group other groups would be doing follow-up activities via the internet or books, practice reading or buddy reading. At the beginning of a session, if shared reading with the whole class occurred, it had a grammar focus using a range of texts (poems, journal stories, and articles and newspaper articles).

This teacher's general strategy was similar to the first teacher's enjoyment of the children through exchanges with humour (discussing *Fantastic Mr. Fox* by Roald Dahl she exclaims: "Wow, it's like Kentucky Fried Chicken isn't it?"). She also used extended discussion in which students and teachers focused on meanings with little explicit attention to teaching strategies. She too carefully selected texts to draw on background knowledge and balanced that with reading complex unfamiliar texts to the class. The presence of the strategies embedded in text discussion was similar too. In the following example the text doesn't say why a character's hair is fair:

> T: What colour is his hair?
>
> C: [several] White.
>
> T: So what's another word for 'white'?
>
> C: [two children] Fair hair.
>
> T: White?
>
> C: Fair.
>
> T: White would be fair isn't it? How come? Why is his hair white?
>
> C: Because he is a chief.
>
> T: Oh you think a chief the hair will suddenly change to white?
>
> C: No.
>
> T: You're not listening to my question. How do we know, or how can we tell that, or why is his hair fair do you think. Maybe …?
>
> C: He's old.
>
> T: He's old. He's old. Good thinking. Okay and so he was having dreams.

Where she differed somewhat from the first teacher was in a more explicit and frequent emphasis on building the students' awareness through articulating intentions and goal setting. An example is how she clarified the assessments students faced. This teacher explicitly taught cloze exercises,

which has an interesting parallel in the subtest gain scores as the largest gain for this class was in the third subtest, which was the cloze test.

> T: One of your tasks today is to do cloze. I've got some paper here. OK let's read the skeleton with our eyes. We're going over it for the second time just for revision. We've done it, we did it yesterday. Were you here xxx?
>
> C: No.
>
> T: That's the first.… you read the whole thing so you can get an idea of what it is about. What is it about?
>
> C: [one child] Maui.
>
> T: Maui.
>
> C: [one child] going fishing.
>
> T: Going fishing with his brothers. And did he catch anything?
>
> C: [several children] Yes.
>
> T: Now tell me, which word should go in the first place?
>
> C: [a few children raise their hands]
>
> T: You have to tell me how did you know? C?
>
> C: Um "morning".
>
> T: Can you read these two sentences to me.
>
> C: Maui had magical powers One [morning] he [went] fishing with [his] brothers.
>
> T: Why is it 'one morning'? Could we have 'one week'? Could we have 'one afternoon'?
>
> C: [several children] Yes.
>
> T: Yes.
>
> T: Cos it makes sense.
>
> This analysis of a paragraph goes for 33 more turns.

Another example occurred in discussing a goal to become better readers through reading at home to practise. The example also illustrates how this classroom functioned as a community of readers. In answer to a question about the goal:

> C: Take your um reader home every night and read it.
>
> T: You should. How many of you are doing that already?
>
> C: [a few children raise their hands]
>
> T: Are you changing your books regularly?
>
> C: [a few children] Yes

T: Are you sharing it with Mum and Dad?

C: [children] Yes.

T: Whose Mum is reading with you? Or some family member?

C: [a few children raise their hands]

T: Are they enjoying the book as much as you're enjoying?

C: [children say] Yes.

T: I wish someone could read with me and read to me but I'm just so busy. I've got books looking at me.… as I've got a little bit more time I'm gonna try reading too. When I was as old as you, I really read my books. And I loved listening to my teacher read them. Who doesn't enjoy reading? We all do don't we?

C: [children say] Yes.

T: Well that's our goal. By the end of the year we have to read …?

C: [children say things]

T: Above our chronological age. Remember it's hard to keep up because every day we are getting …?

Discussion continues for 11 turns

Vocabulary was also clearly a focus seen in interactions which involved explicitly discussing words or phrases such as "again and again and again". As with the first teacher she read high-interest, complex texts, in the early transcript this was *Fantastic Mr. Fox* by Roald Dahl. The text itself contained complex words and language uses.

C: The farmers were determined to dig them out.

T: Very good word. What's the word she used?

C: [a few children] Determined

This teacher also made connections between reading and writing. After reading *Fantastic Mr Fox* she said: "Excellent sentences aren't they which we can use in our writing". Before guided reading she asked students to go over learning intentions on a whiteboard about what the students should be learning from the text.

C: [reading from the whiteboard] Collect words and use them in daily writing.

T: Is that important?

C: [a few children] Yes.

T: When we are doing our writing sometimes we are stuck for words. Yes?

C: [a few children] Yes.

T: We know what we want to say but we can't get the words so where is a good place to find good words to use in our writing?

C: In stories.

T: In reading. In the reading of course. OK we have to look carefully? Second one? [points to the whiteboard]

C: [children read aloud] Expand and elaborate ideas.

T: So in our writing I want to see, I want to see that in your writing you know how we can elaborate ideas. We'll look at how authors write. Not today. Maybe tomorrow OK?

Comparisons between teachers

In both classrooms students made gains over a year; in both these were substantially above normally expected rates of gains. The first teacher represented very high gains and the second gains that were average for the whole group of teachers. The teachers were very similar in many respects. There is very little evidence in the quantitative or the qualitative data to draw distinctions. The one dimension that appears in the quantitative and qualitative data is a sense that the second teacher was more explicit and overt than the first, especially to do with building students' awareness of what was required in their knowledge and performance and checking both by her and requiring the students to check. Overall, the second teacher was more dominant in instruction than the first. These differences are shown quantitatively by differences in the exchange rate for these teachers in the second year as shown in Table 29.

Table 29 **Frequency of exchanges for Teacher 1 and Teacher 2**

	Teacher 1 (high gain)		Teacher 2 (average gain)	
Total exchanges	**Beginning Year 2 33 mins**	**End Year 2 46 mins**	**Beginning Year 2 40 mins**	**End Year 2 49 mins**
Text related	28	43	49	55
Vocabulary questions T	3	12	9	13
Vocabulary comment T	3	4	5	9
Extended talk T	15	30	20	31
Extended talk C	15	26	20	29
Text check T	20	6	35	23
Text check C	11	3	29	19
Incorporation	1	14	3	7
Awareness strategy	10	1	10	0
Awareness other	25	26	69	39
Feedback high	25	44	35	39

4. Discussion

What about tomorrow?

We began this report establishing the project as a systematic replication of a previous intervention which had taken place with seven Mangere schools. The contexts and the theoretical rationales were the same and for that reason we repeated descriptions of the historical and social context. Consistent with the first intervention we return to the general background to locate the significance of the results reported here. The present report adds to our understanding of these contexts. But the present report is also a planned replication and needs to be seen as providing the necessary check on both interventions in the purely scientific sense of checking whether the findings could be repeated. In a sense this is a separate question and findings of similarities and differences add to our judgment of robustness as well as providing finer-grained analyses of how the interventions work in terms of processes and outcomes.

The short history of educational research into the schools in South Auckland provided in the introduction to this report suggested that Māori and Pasifika children in these decile 1 schools were likely to be at risk of low achievement. A landmark study in 1981 proclaimed that "tomorrow may be too late" for the children and their schools (Ramsay et al., 1981).

There has been little evidence of gains in achievement in literacy since the Ramsay report until the NEMP evidence in 2001 (Flockton & Crooks, 2002). But this and the next cycle of national assessments in 2004 (Crooks & Flockton, 2005) contained mixed messages—that although levels of fluency and accuracy of decoding had increased for Māori and Pasifika children, comprehension levels at Years 4 and Years 9 were still low compared with other children and the gaps may have been increasing.

This project set out to ask three general questions.

- Could the processes and findings of the previous intervention (McNaughton et al., 2006) be replicated?
- Can a research–practice collaboration develop cluster-wide and school-based professional learning communities that are able to critically analyse and problem solve issues of instructional effectiveness thereby developing more effective instruction that has a powerful educationally significant impact on Māori and Pasifika children's comprehension at Years 4–8 in seven decile 1 schools?

- Can a set of effective instructional activities be identified that are able to be used by teachers to enhance the teaching of comprehension for Māori and Pasifika children in Years 4–8 in decile 1 schools?

Educationally significant impact?

The answer to the questions about achievement and effective teaching is that it is possible to develop more effective teaching that impacts directly on the reading comprehension achievement of Year 4–8 children. This question can be examined longitudinally with children present at all six data points over three years. This involved three cohorts (Year 4, Year 5 and Year 6) in the Mangere study and one cohort (Year 4) in the present study. As with the previous study the level of gains overall was substantial. The gains in both projects were in the order of nearly one year's gain in addition to nationally expected progress over the thee years of the project (Mangere gain = 0.97 stanine and Otara = 0.85 stanine).

Children who had been at school continuously for three years made these gains with associated effect sizes of d = 0.62 (Mangere) and d = 0.64 (Otara). These effect sizes represent over half a standard deviation difference between the groups' achievement distributions at the beginning and at the end of the projects. The rate of gain in both clusters, illustrated in the best fit regression lines (see Figure 28) was remarkably similar.

A second way of examining the effects is to consider those children present at the beginning and the end of each year (but not necessarily present over three years). Again increases were associated with each year, varying between 0.81 stanine and 0.28 stanine, so that by the third year the total children present at the beginning and end gained 0.39 stanine in Mangere and 0.28 stanine in Otara. Even when considering all the children present from the beginning to the end, including children who subsequently left and those who subsequently entered the school either from earlier levels or as new students from other schools, the levels of achievement at the schools have increased considerably. In Mangere they shifted from 3.1 stanine at the beginning of 2003 (Time 1) to 3.75 stanine at the end of 2005 (Time 6), and in Otara they shifted from 3.02 stanine at the beginning of 2004 (Time 1) through to 3.67 at the end of 2006 (Time 6).

We examine further below the theoretical significance of these findings. It is worth underlining here, however, what the educational significance is. The effectiveness of the teaching increased substantially. At the beginning, teaching was associated with students making expected gains across year levels in their reading comprehension. Unfortunately these students needed to make accelerated gains, because on average they were almost two years behind in achievement levels. The teachers were able to do this. At the end the students were now less than a year behind average national levels. But more importantly, 69 percent of the children (Mangere) and 66 percent (Otara) of the children were now in middle to upper bands of reading comprehension for age level compared with only 40 percent (Mangere) and 34 percent (Otara) at the start. 77 percent of children would be expected in the Average or Above Average bands.

This can be put in a more general educational context. The most general and well-documented acceleration programme we have in New Zealand is associated with gains to middle levels of reading for the classroom for low-progress children after the first year (McDowell, Boyd, & Hodgen, 2005). It achieves this on a daily one-to-one basis for half an hour over 15 to 20 weeks. These are 6-year-olds and the target is the middle band for the school, not national averages. The teaching in the programme reported here occurred in the typical classroom reading sessions, with classes of 20 to 30 students and using the usual instructional vehicles of instruction, which include well-known approaches such as guided reading and deliberate acts of teaching (Ministry of Education, 2006a). More generally, in the United States Borman (2005) shows that national reforms in schools to boost the achievement of children in low-performing schools serving the poorest communities have produced small gains in the short term (of the order of effect sizes of less than 0.20) but that after seven years, in those few schools that sustain reforms over a long period, the effects increase (estimated to be around effect sizes of 0.50). When considered across the United States, while some achievement gains have occurred, they have typically been low and need to be accumulated over long periods of time. Given the longstanding and seemingly intractable nature of the challenge of teaching more effectively in the South Auckland schools, this marks a major breakthrough in our demonstration and understanding of effective teaching.

The quasiexperimental design, with its additional checks through the lagged comparisons with a similar but untreated cluster, and through testing the contribution of subject attrition, give us considerable confidence in attributing these outcomes to the research and development programme. The confidence is increased markedly by the replication of general outcomes. The design is not fully experimental, hence we cannot say without qualification that the research and development programme caused these results. But the inbuilt comparisons against projected levels, the replications within and now across studies, and the patterns of change over time make it highly likely.

The research and development programme involved several components which were added together sequentially. In the following sections we discuss the contribution of different components of the intervention in both studies.

The three-phase model

Research reviews argue that effective educational interventions in general have a component which involves the collective use of evidence, typically from student achievement data, to guide and evaluate teaching practices. Similarly, effective educational interventions have a focus on instructional practices for specific aspects of student learning. The collaborative research and development programme implemented in both the present and the previous study involved a focused intervention incorporating both elements of effective educational interventions in a cluster of poor urban schools with communities that are both culturally and linguistically diverse. Collaboration in the first year (Phase One) entailed the development of a professional learning

community, focused on collecting, analysing, and critically discussing evidence. In the second year (Phase Two) the professional development programme focused on specific aspects of the teaching of reading comprehension. Unlike some other interventions, the specific practices were highly contextualised, developed directly from the profiles of teaching and learning identified in the first phase. Phase Three involved the critical discussion of Phase One and the teaching targeted in Phase Two. Further professional development did not occur, but further components designed to build the critical discussion around evidence through the professional learning communities within and across schools were added.

When considered longitudinally (cohorts over three years) large gains in achievement were associated with Phase One, albeit slightly larger in Otara (Otara = 0.54 stanine) than Mangere (Mangere = 0.47 stanine). Almost identical gains continued in achievement in the second phase, demonstrating the potential for sustaining over the longer term (Mangere = 0.35 stanine; Otara = 0.34 stanine). However, the rates of gain were both smaller and more variable than in the first year, albeit the effect sizes in the second year were comparable to those reported internationally for effective educational interventions (Annan & Robinson, 2005). The rates of gain in the third year were markedly less similar (Mangere = 0.51 stanine; Otara = 0.19 stanine). The only change to this pattern when considering the data in terms of gains in each year for all students present across the time points was the presence of a relatively low rate of gain in Mangere in the second phase.

What was the role of the first phase, the critical analysis and discussion of evidence? The sequence adopted in the educational intervention was one in which critical discussion preceded fine-tuning of instructional practices. As the content for fine-tuning instructional practice was deliberately based on the findings from the critical analysis of evidence, it was not possible in this research design to sequence differently. Nevertheless, given the sequence, it appears that thinking about and critically discussing the evidence at a classroom, school and cluster level provided a significant part of the overall gains in achievement

This finding is consistent with other studies where the critical analysis of data has been linked to sustaining gains in achievement or improving achievement (Phillips et al., 2004; Timperley et al., 2003). This is also consistent with Hawley and Valli's (1999) review of professional development in which they identify critical analysis as a more effective form of professional development than traditional workshop models.

What this finding in turn suggests is that the professional learning communities had the capacity to use the evidence to make changes to existing practices prior to professional development aimed at those practices; however they needed support from researchers to identify the locus of the changes. This is consistent with a view of teachers as having professional expertise and suggests that when teachers engage in problem solving and theorising about their own practices in professional learning communities, the expertise distributed through the community contributes to marked learning (Robinson & Lai, 2006).

However this does not imply that professional development aimed at identifying and fine-tuning specific practices was not needed. Despite the substantial gains in the first phase, they were not sufficient to achieve the goal which each of the school communities has set of parity with national distributions. Moreover, there were cohorts which made the same or even higher gains in the second phase and were only then approaching national levels in their classrooms. So it does not appear that the professional development focused on specific instructional practices was of lesser significance *per se*. Indeed one interpretation of the results is that gains following, or in addition to, analysis are harder to achieve.

There are issues in these findings around the effectiveness of the professional development in the second phase and increasing its effectiveness. The lower gains may have been due to the issue of guaranteeing the fidelity of the programme, which other writers have noted (Newman et al., 2001). However the measures of instruction suggest that teachers fine-tuned their practices in accord with the professional development emphasis. We established the integrity of the treatment by taking samples of classroom teaching. In general the instruction changed in ways that were consistent with the feedback and the instructional hypotheses. This adds confidence to our conclusions about the nature of effective instruction for reading comprehension in this context (see below). It is also possible that the solutions determined collectively in the first phase were incomplete, although many of the dimensions were similar to attributes of effective teaching others have identified (Pressley, 2001). But the professional development in the second phase was associated with additional large gains in some cohorts, indicating the usefulness of the content.

One feature of the programme in the second year in the Mangere schools that was likely to be influential was the variability in engagement of teachers in the professional development programme and delivery of the curriculum associated with the schools. This was not present in the first phase, suggesting that the effectiveness of the professional development around instructional practices was determined by attributes of the schools. One school was clearly less effective in the second phase. Most classes made small gains or actually reduced in stanine averages in the second phase while they had made substantial gains in the first phase. The school results are likely to have been a consequence of its decision to withdraw all but three teachers from the professional development component in the second phase. This indicates that the improvements in achievement attained through critical analysis could not be sustained without continued involvement in the second phase focusing on identifying and fine-tuning specific practices. This sort of variability was not present in the Otara schools and their rates of gain in the second year were consistently higher than the Mangere schools.

The latter finding further implicates the importance of continuing effective leadership in schools and an effective professional learning community as highlighted by previous research (Timperley et al., 2003). Similarly, the results of the third phase also support the significance of the professional learning communities. The third phase deliberately added components to build the sharing of evidence of effective practice. For Mangere the gain for the longitudinal cohorts was the same in the third phase as the first phase, and noticeably larger than the second phase.

However for the Otara schools the gains in the third phase were slightly smaller than those of the second phase, which was smaller than the first phase.

The sustainability phase

The sustainability of school interventions has been identified as a major problem in the research literature (Coburn, 2003). The consensus is that sustaining high-quality interventions is dependent on the degree to which a professional learning community is able to develop (Toole & Seashore, 2002). A learning community can effectively change teacher beliefs and practices through collectively inquiring into their own practices (Annan et al., 2003; Hawley & Valli, 1999; Timperley & Robinson, 2001). We predicted that across the clusters gains would continue to be made given that a professional learning community developed which critically discussed evidence and used that evidence to monitor and modify practices. We had hypothesised that attributes of these communities included being well-versed theoretically, being evidential, being analytic, and being culturally located (that is, locating their knowledge of teaching in learning in the local patterns, including knowing about the strengths and resources of the students and their communities).

Our indicators for these attributes in the third phase included the continued involvement of schools in the process of critical discussion, and the designing, implementing, and collective reporting of classroom-based projects in the teacher-led conference. In general, there was a varied rate of engagement by teachers as well as leaders in the conference for those that participated. The topics for projects were theoretically based, the teachers gathered and reported on evidence, they adopted an analytic stance to that evidence and they related their analyses to the patterns of student learning and teaching in their classrooms. The evidence from the achievement data is that the intervention was sustained at a lower rate in the third year. The indication suggests that involvement in the preparation and presentation of the projects was associated with continued gains.

This study adds to a growing body of research (e.g., Taylor et al., 2005; Timperley et al., 2003) which suggest the importance of promoting the critical analysis of evidence in schools. Further details on how such a process can be developed and what it should look like are given by Robinson and Lai (2006), who present the methodology underpinning the critical analysis used in this study and provide detailed descriptions of the analysis process. This process includes a close examination of students' strengths and weaknesses, and of current instruction, to understand learning and teaching needs. Valid inferences are then drawn from the information through raising competing theories of the "problem" and evaluating the evidence for these competing theories using standards of accuracy, effectiveness, coherence, and improvability.

Our study, however, reveals that there are some conditions to consider when encouraging such analysis in schools. Firstly the findings suggest, as others have found, that such analysis is likely to be dependent on external support in the form of collaborative research–practice–policy

partnerships (e.g., Annan & Robinson, 2005; Lai et al., 2004). We need to consider how to foster such partnerships, both in terms of the kind of partnerships being developed and the infrastructure to support their development and sustainability. (See Annan and Robinson, 2005, for discussion on effective policy–research–school partnerships). Such infrastructure could be in the form of short-term projects such as the Teaching Learning Research Initiative, which is contestable funding by central government for short-term projects that must involve partnerships between schools and researchers. Or it could be longer-term collaborations such as the Woolf Fisher Research Centre, an independent centre developed to improve student learning in a low socioeconomic community through research–school and policy partnerships. Neither are mutually exclusive and policy makers need to consider how best to use different vehicles to achieve their goals.

Secondly, our findings also caution against focusing on critically analysing evidence as a substitute for professional development that focuses on fine-tuning teachers' pedagogical and content knowledge. The data suggest that a successful professional development programme may need to incorporate both elements so that the critical analysis of evidence reveals the students' learning needs, and consequently reveal the content and pedagogical knowledge required to address those needs. This is important as there is a danger, given recent emphasis on analysing data, that we downplay the importance of teachers' understanding of how to teach. Analysing evidence reveals the problem but if teachers do not have the knowledge to know how to address the problem, the impact on student learning outcomes is limited. Conversely, developing specific content knowledge without knowing whether the content being developed matches the needs of the students is also less effective (e.g., Buly & Valencia, 2002).

Thirdly, the complexities in our data around the first three phases of professional development highlight the need for more research to better understand the locus of change in student outcomes and the impact of schools and teachers on those changes. Most research programmes (e.g., Taylor et al., 2005) utilise some combination of critical analysis and fine-tuning of instructional practices within a variety of research–policy–practice partnerships. Far less is known about how various components of these combinations work together to explain the results with different cohorts and in different school contexts. The complexity of our findings on the locus of change suggests that we can enhance our impact if we better understand these complexities and their impact on achievement. Policymakers need to encourage research which collects more detailed data on features of schools and cohorts that may enhance the impact of professional development.

Conclusions

The significance of research-based evidence to inform educational policy and practice has been a major theme in recent commentaries on improving outcomes for children (McCardle & Chhabra, 2004), and especially in the case of children with cultural and linguistic identities associated with minority status in poorer schools (Alton-Lee, 2003). While searching for an evidence base for

effective reading instruction is important, it is also important to demonstrate that the use of that evidence can make a difference and to understand the mediation processes in the use of that evidence.

In this study data on levels of achievement and students' comprehension were collected across age levels and across the cluster of seven schools. In addition, observations of classroom practice provided details of current patterns of instruction. These two sources of evidence were fed back to school leaders and classroom teachers who, with the research team then systematically analysed and developed hypotheses about teaching and learning needs. This process was established in the first phase, continued in the second and was augmented further in the third. This process of critical discussion and analysis of data within the school cluster was based on previous research suggesting that the critical examination of practice in collaborative groups can be effective in creating meaningful and sustainable changes in practice (e.g., Ball & Cohen, 1999; Timperley, 2003; Toole & Seashore, 2002).

The outcomes show that gathering systematic profiles of children's achievement (McNaughton et al., 2006) and of classroom instruction provide one important mechanism for problem solving. The latter adds importantly to our understanding. Patterns in the children's data can be married with patterns in the classroom instruction. For example, without the classroom observation data the patterns of errors in the cloze tests might have suggested the need for more explicit teaching of comprehension strategies (Pressley, 2002). However, the observations revealed that explicit teaching was generally present and occupied significant amounts of teaching time. Rather, the issue was more directly a problem in the purpose of using strategies, that is, constructing meaning from and enjoyment of texts, and specifically the need to use evidence within texts to support those purposes. There are some anecdotal references to this potentially being a problem in strategy instruction (Dewitz & Dewitz, 2003; Moats, 2004), but specific observations for this context were needed.

An interesting feature of the school analysis is that in both studies there were few consistent differences in gains associated with overall initial achievement levels in schools. It might be expected that schools with initially higher achievement gains would benefit more from the analysis and feedback process, analogous to Matthew effects for individuals and groups (Bandura, 1995; Stanovich, 1986). Conversely it might be expected that schools with lower achievement levels would make more gains because of a higher "ceiling", meaning it would be easier to achieve some shifts where the body of student achievement was very low. The absence of these effects suggests that the processes of analysis and feedback were quite robust across schools.

There are several interesting similarities and differences within the general outcomes that warrant further examination. One is that in both the Mangere and Otara data there is evidence for relatively low gains at the Year 6 level. In the analyses of gains in each year for the same year level repeated over the three years it is the Year 6 students who make the lowest gain. This was particularly marked for the Otara results. It is not apparent what this might reflect. There is some evidence of a plateau in reading achievement in the upper primary school from Year 6 through to

Year 8 provided by cross-sectional data from the AsTTle normative profile (Ministry of Education, 2006b). But, unlike the AsTTle results, the low gains were limited to just Year 6.

A difference appeared in the profile of reading comprehension provided by the PAT. Although the overall achievement levels were similar, the children in Otara generally had lower scores in inferential questions compared with the factual questions. There was also a smaller percentage of Otara children in the middle to upper stanine bands of achievement (notably in the bands 7–9). These differences may reflect differences in the schools, in the children, or in the instruction. We can not separate these out, but in the section following, which discusses the observational data, there are some areas of difference which may relate to lower achievement in the upper bands and the need to boost "deeper" levels of comprehension such as inferencing and critical discussion further.

Reading comprehension and effective teaching

The initial profile of student's comprehension in the cluster of decile 1 schools confirmed previous descriptions of "below average" levels in the middle to upper primary school years (Flockton & Crooks, 2001; Crooks & Flockton, 2005; Hattie, 2002) and replicated the findings of the Mangere study (McNaughton et al., 2006). The profile was the same across age levels. However, it is important to note the presence of variability within the profile.

What can be determined from the patterns within and across tests? In the Mangere study we disconfirmed the hypothesis that students had not developed fast and accurate decoding skills known to be a necessary but not sufficient condition for effective comprehending in conventional school texts (Tan & Nicholson, 1997; Pressley, 2002). The findings suggested that widespread problems with decoding skills were unlikely to be the underlying reason for the low STAR results and this appeared to be true for this second cluster. At every year level, the subtest "word recognition" in STAR was higher than any of the other subtests. On average, students got between 60 percent and 80 percent of the words correct, indicating ability to identify words reasonably accurately under timed conditions. These means are between 1.1 and 2 raw scores different from the means reported in the manual for the nationwide sample, indicating that word recognition skills were similar to the country as whole. (Elley, 2001, states that only a raw score difference of 3 to 4 points can be considered significant).

Given this initial pattern, what do the analyses of classroom instruction over the course of the intervention suggest about effective instruction? Other interventions, which have components of collaborative problem solving and fine-tuning of practices based on expert use of evidence, and which are theory driven, have demonstrated gains in reading comprehension (Taylor et al., 2005). In the Mangere study we showed too that an intervention with these components can be effective in raising levels of achievement (McNaughton et al., 2004).

However, the position we have adopted is that while general relationships between instruction and what students learned over the course of the intervention could be assumed, there would be

specific relationships and needs for this particular context. Close examination of these relationships contributes to the twin challenges of applying research-based knowledge to school practices (Pressley, 2002) and the need to continue to tease out attributes of effective instruction with diverse students (Sweet & Snow, 2003).

What we have found in part confirms both what we found in the Mangere study and an already substantial body of generalisable findings. For example, word knowledge can be increased and extended through specific attributes of instruction. Experimental work demonstrates that instruction embedded in texts which provides elaborations of word meanings and repeated exposure to and use of these increase acquisition of targeted words, and there is some evidence too for generalised effects of specific vocabulary instruction (Penno, Wilkinson, & Moore, 2002). In the present study general increases in exchanges which targeted new or unfamiliar words in texts, and which involved extended discussion between teachers and students were associated with increases in vocabulary knowledge on the standardised test.

What the Mangere study and the present study add is that this relationship can be employed in a multicomponent programme of teacher change and also achieve detectable generalisable effects. In the Mangere study the most effective teacher used her selection of texts and specific attributes of interactions to achieve these effects. Importantly, what appeared to distinguish her from a less effective teacher was her own use of a wide and complex vocabulary as well as her expectations that this complexity was appropriate for her students. In the present study vocabulary subtest scores increased (and had the second highest effect size) associated with changes in teacher extended talk and students' extended talk, much of it introducing and elaborating vocabulary. An additional analysis revealed that teachers increased their explicit teaching of words other than Technical or Topic Related words to about four words per session in addition to the existing rates of explicitly teaching Technical and Topic Related words each on average once per session. A parallel can be drawn with early language development, where quality of language use, such as complexity and type of vocabulary, is important. Frequency or repetition of language use also need to be considered (Hart & Risley, 1995). We did not conduct extra analyses of the Mangere teachers' explicit teaching of vocabulary so we cannot compare rates, but it is interesting to note that the increase in the vocabulary subtest scores in Otara was lower than the Mangere gains in vocabulary. A fruitful area for further research is examining optimal rates and forms of vocabulary teaching with these students. It is also interesting to note that the overall rate of teacher interaction was higher in this replication than in the original (Mangere) study, possibly creating a teacher-dominance effect and putting a ceiling on achievement gains at the upper stanine levels.

A solid research base exists which provides considerable evidence for the significance of developing reading strategies (Pressley, 2002). But we found in the Mangere study that there was a specific problem with strategy instruction and we found it again here. It was a problem of continued explicit strategy instruction as an end in itself, being deployed in a formulaic way as routines to be run off rather than as strategic acts whose use and properties are determined by the overarching goal to enable readers to construct and use appropriate meanings from texts (Baker,

2002; Moats, 2004). Several of the quantitative measures were indicators of the change in this problem between our two studies. One was explicit instruction to build students' awareness and another was checking evidence by both teachers and students. In Otara, teachers generally reduced their explicit focus on teaching strategies; overall it may have been too much because the gains in paragraph comprehension over two years were not as much as the gains in the Mangere study. The increased focus on checking over the intervention was associated with the gains in component tests including paragraph comprehension. Our hypothesis is that maintaining the focus on using texts to clarify, confirm, or resolve meanings and avoiding the risk of making strategies ends in themselves may be particularly important to the continued effectiveness of strategy instruction.

As we noted in the Mangere study, the solution to this risk lies in the collective evidence-based problem solving and the increased knowledge teachers developed to understand the nature of comprehending, learning and teaching, and characteristics of effective teaching. These are features of effective programmes that have also been identified by other researchers (Taylor et al., 2005). In addition to what we said about this in the Mangere study, there is a need for ongoing monitoring of classroom instruction continuing past the initial experimental demonstrations because there is the risk, or at least a possibility as noted above, that too much change in one direction may occur.

A body of evidence demonstrates that effective comprehension of school texts and effective learning from school texts is dependent on the learner developing awareness, in the sense of monitoring of and control over performance (Guthrie & Wigfield, 2000). The generalised significance of this feature for classroom learning, especially for culturally and linguistically diverse students, has been argued by a number of researchers (e.g., McNaughton, 2002). Developing greater awareness of the requirements of literacy instruction and the relationships between current knowledge and these requirements was a component in a previous study of culturally and linguistically diverse students in beginning instruction (Phillips et al., 2004) and it was a component of the Mangere study too. As in those studies, in the present study teachers increased their instruction which directed students to consider goals and intentions and aspects of their knowledge and performance other than specific comprehension strategies. In general, the teachers changed considerably in this area. An indication of this significance was contained in the case studies. Both teachers increased their focus on awareness, but interestingly the second (medium gain) teacher developed very high rates of exchanges focusing on awareness. This might also indicate the tendency to routinise this aspect of teaching beyond when it is needed, or to use it with greater frequency than it might be needed at these levels.

As with the Mangere study, all the teachers provided a high frequency of informative feedback to students. Hattie (1999) has argued that uncritical acceptance of student responses is a feature of New Zealand classrooms and these data appear to support his contention. The data on the teacher questions and checking in the baselines appear to support the prediction from this that students' learning would be enhanced if the adequacy of responses was made clear and grounded in evidence that students could access and check themselves. Feedback might also convey the

expectation of students being able to succeed with difficult tasks, an important component in the development of self-efficacy (Guthrie & Wigfield, 2000).

The concept of guidance carried in discourse patterns is central to our design of effective comprehension instruction (Pressley, 2002). What has been signalled here again is that there is something like a curvilinear relationship between guidance and the effectiveness of instruction. It is possible to have too much just as it is possible to have too little, both at one point in time and over time. The case studies illustrate this, for example around the functions of questioning and the risk of teacher dominance (Cazden, 2001). Certainly, the general concept of planning for dynamic changes in guidance to support growing independence draws attention to the risks of too much guidance. But this understanding needs constant application with teachers monitoring the degree to which their moment by moment interactions do or do not support more complex forms of engagement, and this is related closely to the assumptions that they have about the capabilities of children.

Again like the Mangere data, a relationship that was not well clarified concerns the use of students' cultural and linguistic backgrounds in instruction. Overall a small but noticeable proportion of teachers' exchanges incorporated aspects of students' event knowledge and skills both in the Mangere study and in the present study. Quantified, it was about five exchanges per session. The evidence for effective teaching being very sensitive to and capitalising on backgrounds as resources seems clear, especially in the context of culturally and linguistically diverse students (Alton-Lee, 2004; McNaughton, 2002). This is highly contingent on knowing about students' knowledge and skills and being able to make connections for the students between current knowledge and needed knowledge. We do not know what optimal rates might be, but with the data we have provided here there is a basis for future exploration of this question.

But the relationships are not simple. As with guidance more generally, there are a set of balances here. One is balancing enhancing the match between backgrounds and activities by redesigning activities to incorporate cultural and linguistic resources with developing increased awareness of classroom requirements including the mismatch between current expertise and classroom instruction (Phillips et al., 2004). In the current study it is not apparent what the appropriate balance might be. Whether further gains could have been achieved with increases in this attribute of teaching is not known. What is indicated is that the use of cultural and linguistic resources does not necessarily increase as a function of increasing the range of instructional strategies for reading comprehension *per se*.

Both case study teachers looked very similar in this respect, and both balanced this with developing students' awareness of the goals of classroom activities and formats (including the teachers' own expectations). Both teachers seemed successful in creating bridges between background knowledge and the requirements of classroom activities. They knew their children very well, including the likely areas where decoding could be a problem, and could judge the usefulness of referring to or activating particular event knowledge. They were both sufficiently at ease with the children to engage in humorous exchanges.

The need for exposure to, and extensive practise with, core text-based activities is also well documented (Guthrie & Wigfield, 2000; Pressley, 2002; Stanovich et al., 1996). The descriptions from this study once again highlight this need at various levels of teaching, from the selection and use of a variety of suitable texts and classroom management that maximises engagement with these texts, through to interactions during text-based activities which increase involvement in and learning from the activity. But a further payoff for increased exposure to texts is indicated in these data. The baseline analyses had shown that levels of accuracy of decoding were not a major problem in general for the students. Decoding problems therefore were not targeted in the instruction. What is interesting to note is that despite this, levels of decoding as measured on the standardised test were affected by the instruction focused on text reading. We found this in the Mangere study and we have replicated this here. Other researchers have also found this relationship, where an intervention targeted on comprehension and based on text-based activities had a positive impact on decoding (Lefevre et al., 2003).

These descriptions contribute to meeting educational research and development needs, identified by Pressley (2002) as the need to apply knowledge to teaching contexts, and by Sweet and Snow (2003) as the need to fill gaps in our research knowledge. In our view, to successfully apply knowledge to teaching contexts, instructional principles need to be designed to fit context-specific needs. These needs can be determined from past histories of schooling and contemporary profiles. The descriptions provided here contribute to the research problem by supporting and extending our understanding of basic attributes of effective teaching of reading comprehension.

References

Alton-Lee, A. (2003). *Impact of teachers and schools on variance in outcome.* Unpublished paper, Ministry of Education, Wellington.

Alton-Lee, A. (2004). *A collaborative knowledge building strategy to improve educational policy and practice: Work in progress in the Ministry of Education's Iterative Best Evidence Synthesis Programme.* Paper presented for a symposium at the annual conference of the New Zealand Association for Research in Education, Wellington, 25 November 2004.

Annan, B. (1999). *Strengthening education in Mangere and Otara. Summary of the SEMO annual report: The evolution of a 3-way partnership, schooling and development project.* Wellington: Ministry of Education.

Annan, B., & Robinson, V. (2005, April). *Improving learning processes for practitioners involved in school reforms.* Paper presented at the American Educational Research Association conference, 11–15 April, Montreal, Canada.

Annan, B., Lai, M. K., & Robinson, V. M. J. (2003). Teacher talk to improve teacher practices. ***set:*** *Research information for teachers,1,* 31–35.

Awatere-Huata, D. (2002). *The reading race. How every child can learn to read.* Wellington: Huia Publishers.

Baker, L. (2002). Metacognition in comprehension instruction. In C. C. Block and M. Pressley (Eds.), *Comprehension instruction: Research-based best practices* (pp. 7–95). New York: Guilford Press.

Ball, D., & Cohen, D. (1999). Developing practice, developing practitioners: Toward a practice based theory of professional education. In G. Sykes & L. Darling-Hammond (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 3–32). San Francisco: Jossey-Bass.

Bandura, A. (Ed.). (1995). *Self-efficacy in changing societies.* New York: Cambridge University Press.

Biemiller, A. (1999). *Language and reading success.* Cambridge, MA: Brookline Books.

Biemiller, A. (2001). Teaching vocabulary: Early, direct and sequential. *American Educator,* Spring. Retrieved 2005, from http://www.aft.org/pubs-reports/american_educator/issues/spring01/index.htm

Bishop, R. (2004). *Te kotahitanga.* Retrieved 2005**,** from www.minedu.govt.nz/goto/tekotahitanga

Block, C. C., & Pressley, M. (Eds.). (2002). *Comprehension instruction: Research-based best practices.* New York: Guilford Press.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher*, *33*(8), 3–15.

Borman, G. D. (2005). National efforts to bring reform to scale in high-poverty schools: Outcomes and implications. In L. Parker (Ed.), *Review of research in education* (Vol. 29, pp. 1–27). Washington DC: American Educational Research Association.

Brown, A. L. (1997). Transforming schools into communities of thinking and learning about serious matters. *American Psychologist*, *52*(4), 399–413.

Bruno, J. E., & Isken, J. A. (1996). Inter-intraschool site student transiency: Practical and theoretical implications for instructional continuity at inner city schools. *Journal of Research and Development in Education*, *29*(4), 239–252.

Buly, M. R., & Valencia, B. W. (2002). Below the bar: Profiles of students who fail state reading assessments. *Educational Evaluation and Policy Analysis*, *24*(3), 219–239.

Cazden, C. (2001). *Classroom discourse* (2nd ed.). Portsmouth, NH: Heinemann.

Chatterji, M. (2005). Achievement gaps and correlates of early mathematics achievement: Evidence from the ECLS K–first grade sample. *Education Policy Analysis Archives*, *13*(45). Retrieved 2005 from http://epaa.asu.edu/epaa/v13n46/v13n46.pdf

Clay, M. M. (2002). *An observation survey of early literacy achievement* (2nd ed.). Auckland: Heinneman.

Coburn, C. E. (2003). Rethinking scale: Moving beyond numbers to deep and lasting change. *Educational Researcher*, *32*(6), 3–12.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum & Associates.

Crooks, T., & Flockton, L. (2005). *Reading and speaking assessment results 2004* (National Education Monitoring Report 34). Dunedin: Educational Assessment Research Unit.

Currie, G. A. (1962). *The report of the New Zealand commission on education in New Zealand.* Wellington: Government Printer.

Darling-Hammond, L., & Bransford, J. (Eds.). (2005). *Preparing teachers for a changing world.* San Francisco: John Wiley.

Delpit, L. (2003). Educators as "Seed people" growing a new future. *Educational Researcher*, *32*(7), 14–21.

Department of Education. (1930). *New Zealand education gazette*. Wellington: Author.

Dewitz, P., & Dewitz, P. K. (2003). They can read the words but they can't understand: Refining comprehension assessment. *The Reading Teacher*, *56*(5), 422–435.

Dickinson, D. K., & Tabors, P. O. (2001). *Beginning literacy and language: Young children learning at home and school.* Baltimore, MD: Paul Brookes Publishing.

Dyson, A. H. (1999). Transforming transfer: Unruly students, contrary texts and the persistence of the pedagogical order. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education, 24* (pp. 141–172). Washington DC: American Educational Research Association.

Education Gazette. (1930, December). Survey of native schools. *The New Zealand Education Gazette, 9*(12) 248.

Elley, W. (1992). *How in the world do children read?* New York: International Association for the Evaluation of Educational Achievement.

Elley, W. (2001). *STAR Supplementary test of achievement in reading: Years 4–6.* Wellington: New Zealand Council for Educational Research.

Elley, W. (2005). On the remarkable stability of student achievement standards over time. *New Zealand Journal of Educational Studies*, *40*(1 & 2), 3–23.

Elley, W. B., & Croft, A. C. (1989). *Assessing the difficulty of reading materials: The noun frequency method* (Rev. ed.). Wellington: New Zealand Council for Educational Research.

Flockton, L. (2003). *Nationally speaking: Examining the NEMP data*. Keynote address to the Learning Media National Literacy Symposium, Wellington, 19–20 June 2003.

Flockton, L., & Crooks, T. (2001). *Reading and speaking: Assessment results 2000* (National Education Monitoring Report 19). Dunedin: Otago University for the Ministry of Education.

Flockton, L., & Crooks, T. (2002). *Writing assessment results 2002* (National Education Monitoring Report 27). Wellington: Ministry of Education.

Garcia, G. E. (Ed.). (2003). *The reading comprehension development and instruction of English-language learners.* New York: Guilford Press.

Guthrie, J. T., & Wigfield, A. (2000). Engagement and motivation in reading. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of Reading Research: Volume III* (pp. 403–422). Mahwah, NJ: Lawrence Erlbaum & Associates.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children.* Baltimore, MD: Paul Brookes.

Hattie, J. (1999, August). *Influences on student learning: Inaugural lecture: Professor of Education, University of Auckland, 2 August 1999*. Retrieved 15 April, 2008, from The University of Auckland, Faculty of Education Web site: www.education.auckland.ac.nz/uoa/fms/default/education/staff/Prof.%20John%20Hattie/Documents/Presentations/influences/Influences_on_student_learning.pdf

Hattie, J. (2002). *What are the attributes of excellent teachers?* Paper presented at the New Zealand Council for Educational Research conference, Wellington, October 2002.

Hawley, W. D., & Valli, L. (1999). The essentials of effective professional development: A new consensus. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as a learning profession* (pp. 127–150). San Franscisco: Jossey-Bass.

Hunn, J. K. (1961). *Report on the Department of Māori Affairs: with statistical supplement.* Wellington: Government Print.

IEA. (1992). *The reading literacy study.* Retrieved 2005, from http://www.iea.nl/reading_literacy.html

Lai, M. K., McNaughton, S., MacDonald, S., Amituanai-Toloa, M., & Farry, S. (2006). *Replication of a process.* Paper presented at the American Educational Research Association conference, 9–14 April, San Franscisco.

Lai, M. K., McNaughton, S., MacDonald, S., & Farry, S. (2004). Profiling reading comprehension in Mangere schools: A research and development collaboration. *New Zealand Journal of Educational Studies*, *39*(2), 223–240.

Learning Media. (2003). *Good thinking: Comprehension and critical thinking in the classroom.* Learning Media national literacy symposium, Victoria University, Wellington.

Lee, C. D. (2000). Signifying in the zone of proximal development. In C. Lee & P. Smagorinsky (Eds.), *Vygotskian perspectives on literacy research: Constructing meaning through collaborative inquiry* (pp. 191–225). Cambridge: Cambridge University Press.

Lefevre, D. M., Moore, D. W., & Wilkinson, I. A. G. (2003). Tape-assisted reciprocal teaching: Cognitive bootstrapping for poor decoders. *British Journal of Educational Psychology*, *73,* 37–58.

Literacy Experts Group. (1999). *Report to the Secretary for Education.* Wellington: Author.

Literacy Taskforce. (1999). *Report of the Literacy Taskforce.* Wellington: Ministry of Education.

McCall, R. G., & Green, B. L. (2004). Beyond the methodological gold standards of behavioural research: Considerations for practice and policy. *Social Policy Report: Giving Child and Youth Development Knowledge Away, 18,* (2), 3–19.

McCardle, P., & Chhabra, V. (2004). *The voice of evidence in reading research.* Baltimore, MD: Brookes Publishing.

McDowell, S., Boyd, S., & Hodgen, E. (2005). *Evaluation of the effectiveness of Reading Recovery particularly for Māori and Pasifika students (2004–2005).* Wellington: New Zealand Council for Educational Research.

McNaughton, S. (1999). Developmental diversity and literacy instruction over the transition to school. In J. S. Gaffney & B. J. Askew (Eds.), *Stirring the waters: A tribute to Marie Clay* (pp. 3–16). Portsmouth, NH: Heinemann.

McNaughton, S. (2000). *Submission to the Education and Science Committee on 'The Inquiry by the Education of Science Committee into the teaching of reading.*'

McNaughton, S. (2002). *Meeting of minds.* Wellington: Learning Media.

McNaughton, S., Lai, M., MacDonald, S., & Farry, S. (2004). Designing more effective teaching of comprehension in culturally and linguistically diverse classrooms in New Zealand. *Australian Journal of Language and Literacy*, *27*(3), 184–197.

McNaughton, S., & MacDonald, S. (2004). *A quasi-experimental design with cross-sectional and longitudinal features for research-based interventions in educational settings.* Manuscript submitted for publication.

McNaughton, S., MacDonald, S., Amituanai-Toloa, M., Lai, M., & Farry, S. (2006). *Enhanced teaching and learning of comprehension in Years 4–9 in seven Mangere schools.* Retrieved 31 March 2008, from http://www.tlri.org.nz/pdfs/9206_finalreport.pdf

McNaughton, S., Phillips G. E., & MacDonald, S. (2003). Profiling teaching and learning needs in beginning literacy instruction: The case of children in "low decile" schools. *New Zealand Journal of Literacy Research*, *35*(2), 703–730.

Ministry of Education. (2005). *Making a bigger difference for all students. Mangaia he hurahi hei whakarewa ake i nga tauira katoa. Schooling strategy 2005–2010.* Wellington: Ministry of Education.

Ministry of Education. (2006a). *Effective literacy practices in years 5 to 8.* Wellington: Learning Media.

Ministry of Education. (2006b). *Information kit: Student achievement in reading*. Wellington: Author.

Mitchell, L., & Cubey, P. (2003). Characteristics of effective personal development linked to enhanced pedagogy and children's learning in early childhood settings: Best evidence synthesis. Wellington: Ministry of Education.

Moats, L. C. (2004). Science, language, and imagination in the professional development of reading teachers. In P. McCardle & V. Chhabra (Eds.), *The voice of evidence in reading research* (pp. 269–287). Baltimore, MD: Brookes.

Nash, R., & Prochnow, J. (2004). Is it really teachers? An analysis of the discourse of teacher effects on New Zealand educational policy. *New Zealand Journal of Educational Studies*, *39*(2), 175–192.

NEMP. (2004). *Forum comment: 2004 assessment reports: Music–Technology–Reading & speaking.* Retrieved 2005 from www.nemp.otago.ac.nz/forum_comment/

New London Group. (1996). A pedagogy of multiliteracies: Designing social features. *Harvard Educational Review*, *66,* 60–92.

Newman, F. M., Smith, B., Allensworth, E., & Bryk, A. S. (2001). Instructional program coherence: What it is and why it should guide school improvement policy. *Educational Evaluation and Policy Analysis*, *23*(4), 297–321.

Nicholson, T. (2000). *Reading the writing on the wall: Debates, challenges and opportunities in the teaching of reading.* Palmerston North: Dunmore Press.

Openshaw, R., Lee, G., & Lee, H. (1993). *Challenging the myths*. Palmerston North: Dunmore Press.

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly*, *40*(2), 184–202.

Penno, J. F., Wilkinson, I. A. G., & Moore, D. W. (2002). Vocabulary acquisition from teacher explanation and repeated listening to stories: Do they overcome the Matthew Effect? *Journal of Educational Psychology*, *94*(1), 23–33.

Phillips, G., McNaughton, S., & MacDonald, S. (2001). *Picking up the pace: Effective literacy interventions for accelerated progress over the transition into decile 1 schools.* Auckland: Child Literacy Foundation and Woolf Fisher Research Centre.

Phillips, G., McNaughton S., & MacDonald, S. (2004). Managing the mismatch: Enhancing early literacy progress for children with diverse language and cultural identities in mainstream urban schools in New Zealand. *Journal of Educational Psychology*, *96*(2), 309–323.

Pogrow, S. (1998). What is an exemplary program, and why should anyone care? A reaction to Slavin and Klein. *Educational Researcher*, *27*(7), 22–29.

Pressley, M. (2000). What should comprehension instruction be the instruction of? In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of Reading Research, Vol. III* (pp. 545–561). Mahwah NJ: Lawrence Erlbaum & Associates.

Pressley, M. (2001). Comprehension instruction: what makes sense now, what might make sense soon. *Reading Online*, *5*(2). Retrieved 2005, from http://www.readingonline.org/articles/art_index.asp?HREF=handbook/pressley/index.html

Pressley, M. (2002). Comprehension strategies instruction: A turn-of-the-century status report. In C. C. Block & M. Pressley (Eds.), *Comprehension instruction: Research-based best practices* (pp. 11–27). New York: Guilford Publications.

Ramsay, P. D. K., Sneddon, D. G., Grenfell, J., Ford, I. (1981). *Tomorrow may be too late. Final report of the Schools with Special Needs Project.* Hamilton: University of Waikato.

Raudenbusch, S. W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher*, *34*(5), 25–31.

Reid, N. A., & Elley, W. B. (1991). *Revised Progressive Achievement Tests: Reading comprehension.* Wellington: New Zealand Council for Educational Research.

Risley, T. R., & Wolf, M. M. (1973). Strategies for analyzing behavioral change over time. In J. R. Nesselroade & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological issues* (pp. 175–183). New York: Academic Press.

Robinson, V., & Lai, M. K. (2006). *Practitioner research for educators: A guide to improving classrooms and schools.* Thousand Oaks, CA: Corwin Press.

Shadish, W. R., Campbell, D. T., & Cook, T. D. (2002). *Experimental and quasi-experimental designs for research.* Boston: Houghton Mufflin.

Sidman, M. (1960). *Tactics of scientific research.* New York: Basic Books.

Smith, J. W. A., & Elley, W. B. (1994). *Learning to read in New Zealand.* Auckland: Longman Paul.

Snow, C. E., Burns, M. S., & Griffen, P. (1998). *Preventing reading difficulties in young children.* Washington DC: National Academy Press.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360–401.

Stanovich, K. E., West, R. F., Cunningham, A. E., Cipielewski, J., & Siddiqui, S. (1996). The role of inadequate print exposure as a determinant of reading comprehension problems. In C. Cornoldi and J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (pp. 15–32). Mahwah, NJ: Lawrence Erlbaum.

Statistics New Zealand. (2002). *Change in ethnicity question; 2001 census of population and dwellings.* Retrieved 2005, from http//www.stats.govt.nz

Sweet, A. P., & Snow, C. E. (Eds.). (2003). *Rethinking reading comprehension.* New York: Guilford Press.

Tan, A., & Nicholson, T. (1997). Flashcards revisited: Training poor readers to read words faster improves their comprehension of text. *Journal of Educational Psychology*, *89,* 276–288.

Taylor, B. M., Pearson, P. D., Peterson, D., & Rodriguez, M. C. (2005). The CIERA School Change Framework: An evidence-based approach to professional development and school reading improvement. *Reading Research Quarterly*, *40*(1), 40–69.

Taylor, B., Peterson, D., Pearson, D., Janynes, C., Knezek, S., Bender, P., & Sarroub, L. (2001, April). *School reform in reading in high-poverty schools.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, Washington, 2001.

Thomas, G., & Tagg, A. (2005). Evidence for expectations: Findings from the Numeracy Project longitudinal study. In *Findings from the New Zealand Numeracy Development Project 2004* (pp. 35–46). Wellington: Ministry of Education.

Tibble, P. (2002). The plum tree. *School Journal, Part 1*(2), 2–7. Timperley, H. (2003). *Shifting the focus: Achievement information for professional learning.* Wellington: Ministry of Education.

Timperley, H., Phillips, G., & Wiseman, J. (2003). *The sustainability of professional development in literacy parts one and two.* Wellington: Auckland UniServices Ltd for the Ministry of Education.

Timperley, H. S., & Robinson, V. J. M. (2001). Achieving school improvement through challenging and changing teachers' schema. *Journal of Educational Change*, *2,* 281–300.

Toole J. C., & Seashore, L. K. (2002). The role of professional learning communities in international education. In K. Leithwood & P. Hallinger (Eds.), *Second international handbook of educational leadership and administration* (pp. 245–279). Dordrecht, The Netherlands: Kluwer Academic.

Tunmer, W. E., Chapman, J. W., & Prochnow, J. E. (2004). Why the reading achievement gap in New Zealand won't go away: Evidence from the PIRLS 2001 International Study of Reading Achievement. *New Zealand Journal of Educational Studies*, *39*(1 & 2), 255–274.

Utai, B. & Rose, J. (2002). A silent world. *School Journal*, *Part 2*(2), 8–11.

Wagemaker, H. (1992). Preliminary findings of the IEA Reading Literacy Study: New Zealand achievement in the national and international context. *Educational Psychology*, *12,* 195–213.

Whitehurst, G. J., & Lonnigan, C. J. (2001). Emergent literacy: Development from pre-readers to readers. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research* (pp. 11–29). New York: Guilford Press.

Wood, D. (1998). *How children think and learn* (2nd ed.). Oxford: Blackwell.