Sustained Acceleration of Achievement in Reading Comprehension: The New Zealand Experience

Mei Kuin Lai, Stuart McNaughton, Meaola Amituanai-Toloa, Rolf Turner, and Selena Hsiao, University of Auckland*

Schooling improvement interventions for culturally and linguistically diverse students from poorer communities need to solve a set of theoretical challenges relating to more effective literacy instruction. Although recent commentaries have suggested that some of the pressing issues in beginning reading instruction have been resolved, overall effectiveness in teaching reading comprehension is limited, and that research has not had much impact on effective comprehension instruction (Pressley, 2002; Sweet & Snow, 2003).

The need to address effectiveness in teaching reading comprehension is particularly significant for schools serving culturally and linguistically diverse populations (Garcia, 2003). The challenge to meet these needs is pressing in New Zealand where—despite the fact that, on average, students in the middle years of school have high levels of reading comprehension judged by international comparisons—there are large disparities in the distribution of achievement (Alton-Lee, 2004). These disparities are between children from both Māori (indigenous) and Pasifika (immigrants from the Pacific Islands) communities in urban schools with the lowest employment and income levels, and other children. Māori and Pasifika children score lower in reading comprehension measures than children from other ethnic groups. Since at least the 1950s, numerous reports have identified these disparities (e.g., Openshaw, Lee, & Lee, 1993), with one in 1981 calling them a crisis in urgent need of a solution (Ramsay, Sneddon, Grenfell, & Ford, 1981).

Like other countries, New Zealand is concerned with disparities in literacy achievement and has responded to this enduring “education debt” (Ladson-Billings, 2006, p. 3) with programs of schooling improvement and reform at local, district, and even national levels. Since 1998, New Zealand has focused on resolving these disparities. The national policy shifts have led to deployment of resources and to fine-tuning of early literacy programs. These changes have been associated with reductions in the disparities in accuracy and fluency of early reading at a national level (Crooks & Flockton, 2005). Experimental evidence supports the conclusion that specific changes in beginning instruction that have been implemented in groups of schools have been effective (e.g., Phillips, McNaughton, & MacDonald, 2004).


Originally printed in Reading Research Quarterly, 44(1), 30–56. Copyright © 2009 by the International Reading Association.
However, the evidence—some of it presented in this report—also indicates that there has been little, if any, impact on reading comprehension from Year 4 of schooling. Indeed, it appears that the gaps may have increased nationally (Crooks & Flockton, 2005). Although depressing, this is not surprising theoretically. Much of the knowledge and many of the skills required for early fluency and accuracy in reading come from acquiring discrete bodies of knowledge. Paris (2005) called these constrained skills, which he claimed are learned relatively easily. The more language-based and content-dependent nature of comprehension requires unconstrained skills, which are more difficult to both teach and learn. In developmental terms, becoming a good decoder is a necessary but not sufficient condition for good comprehension, and effective instruction for decoding does not necessarily presage later development (McNaughton, 2002).

There has been little evidence that the issue of disparities in literacy achievements between groups can be solved easily in schools. In the United States, Borman (2005) showed that national reforms of schools to boost the achievement of children in low-performing schools serving the poorest communities have produced small gains in the short term, with effect sizes of the order of less than 0.20. For those few schools that sustained reforms over a longer period of around seven years, the effects increased (estimated effect sizes of about 0.50). Borman concluded that although nationally 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, individual studies from the United States have shown that clusters of schools serving minority children have been able to increase the achievement of children in reading comprehension. In one set of studies, Taylor, Pearson, Peterson, and Rodriguez (2005) intervened in high-poverty schools with carefully designed professional development research and development. They reported small cumulative gains across two years.

Implementing effective interventions poses a major challenge in New Zealand schools serving poorer communities with high numbers of Māori and Pasifika students. The goal is not just to produce achievement gains with acceptable effect sizes. The issue is accelerated achievement. Students need to make more than just an expected rate of gain.

This need for acceleration was recognized by the designer of Reading Recovery. Clay’s (1979, 2005) developmental argument was that in order for an early intervention program to be functional for an individual, it needed to change the rate of acquisition to a rate of progress faster than the cohort to whom the individual belonged. This acceleration was needed so that over the brief but intensive period of the individualized intervention, a learner would come to function within the average bands required for his or her classroom. Groups of students from particular cultural groups who have not been well served by school instruction also need to make accelerated gains, to come to function like other students at equivalent levels. Their rate of progress needs to be higher than comparison cohorts. The issue for these students is not the same as in Reading Recovery in that the target is not for a group of students to come to function as a group within
average bands. Rather, in the ideal case, the distribution of students needs to approximate an expected distribution—in the case of New Zealand students, the New Zealand national distribution. The probability of being in the lower (or indeed the upper) “tail” of the distribution should be no more than what would be expected for the population as a whole.

The idea of acceleration in terms of faster-than-normal rates of progress so that the changed distribution comes to match an expected distribution implies several criteria for judging the effectiveness of an intervention. Acceleration needs to be robust in the sense that the rate of gain is increased relative to expected gain across a defined period of time, say over a school or a chronological year. The gain needs to approximate an expected distribution, which likely means the gains need to be sustainable, at least in the medium term, say two to three years. Finally, because of the issue of approximating an expected distribution, there is a need to know that different subgroups within the total group have gained similarly. There is evidence that some interventions and school reforms can have differential effects for different level groups (Correnti, Rowan, & Camburn, 2003; Hubbard, Mehan, & Stein, 2006). Similarly, the ubiquitousness of “Matthew effects” (Stanovich, 1986, p. 381) is well known in educational interventions. Such effects should be avoided in the microcosm of an intervention, where differences among subgroups may become further exacerbated and all groups should make accelerated gains.

The presence of subgroup differences in reading comprehension achievement in mainstream schooling is well documented (Snow, Burns, & Griffin, 1998). In New Zealand, these differences include gender and ethnicity differences as well as differences associated with language background. For example, national monitoring data for 9-year-olds (Crooks & Flockton, 2005) shows that females outperform males on average by an effect size of 0.22 across a range of reading tasks, but most of the differences between males and females are on comprehension tasks. Similarly, the average effect size for the difference between Māori and Anglo-European students across tasks was 0.38, again most of the differences attributable to reading comprehension measures. The effect size favoring students for whom English was the predominant language at home was 0.23. From the general experimental literature, we also know that interventions can produce Matthew effects where those children who know or can do more learn more from a new procedure, an example being instruction designed to boost vocabulary learning (Penno, Wilkinson, & Moore, 2002).

The impact of schooling improvement interventions on such subgroups is not well known (Borman, 2005). There is some evidence that different types of programs are differentially effective with the age or year level of student (Correnti et al., 2003), which might suggest that in a highly prescribed intervention, some students would benefit more than others or that some students would learn less than others. For example, more advanced students might benefit from a program with more advanced instructional elements that allow them to develop unconstrained skills, but they may be limited by a program focused mainly on constrained
skills (Paris, 2005). The intervention we will describe in this research was not initially prescribed but nevertheless was, through a process of development with the schools, highly specific. The process, which led to a controlled fine-tuning of existing instruction, was predicted to be both generic and adaptable enough to serve the needs of the subgroups.

An added problem for schooling improvement interventions designed to produce accelerated gains in achievement over several years is the presence of summer effects, the differential growth in learning over the months when schools are closed (Cooper, Charlton, Valentine, Muhlenbruck, & Borman, 2000; Entwisle, Alexander, & Olson, 1997). Students from poorer communities and minority students make less progress than other students do over this period, contributing to a widening gap in achievement. In Heyns’s (1978) classic study, sixth-grade low-income black students lost almost a quarter of a grade on the word knowledge subtest of the Metropolitan Achievement Test, and lowest income white students made almost no gains. Heyns showed that between half and two thirds of the annual learning gap between white children from high-income homes and black children from the poorest homes accrued during the summer months. The gains over the school year were much closer for all groups.

Although there is no New Zealand study that has taken repeated measures over successive years on which to draw, the extent of the effect captured in reviews suggests that similar summer effects would likely be present. This poses two potential issues for the present study: First, a methodological issue is modeling the growth, taking these assumed effects into account. When achievement is measured at the beginning and end of each academic year, the likely shape of growth over three years is therefore likely to be a staircase of some sort (Borman, 2000). The more substantive issue is the sheer challenge of designing powerful school effects that over time are greater than these assumed summer effects. This is an added challenge to meeting the criteria of effective acceleration.

Can an intervention be designed within the framework of schooling improvement that solves the problem of acceleration meeting criteria for effectiveness? We approached this challenge in the following way. We started with known attributes of effective schooling improvement in reading comprehension that are assumed to apply equally for culturally and linguistically diverse students from poorer communities. Several dimensions have been identified in the school reform literature. They include the form and function of professional learning communities; the specificity of the program; the level and quality of implementation; the relationships between the developer and the local school and school district; the coordination and fit of the model to local circumstances; and the degree to which the program is articulated and elaborated (Borman, 2005; Cohen & Ball, 2007; Datnow & Stringfield, 2000; Raphael, Goldman, Au, & Hirata, 2006). Given the finding that more collaborative programs in literacy focused on teachers’ knowledge and problem solving of their practices in professional learning communities may be more effective in the later years, we argue that the complexity associated with effective instruction for reading comprehension requires much
more localized and informed adaptive expertise (Bransford, Derry, Berliner, & Hammerness, 2005).

Our reading of this literature suggests two major hypotheses for how instruction for culturally and linguistically diverse students in poor urban schools might accelerate achievement gains over a sustained period of time. The first is that instructional practices drawing on evidence about teaching and learning requirements in specific contexts need to be developed, and the second is that professional learning communities need to be able to fine-tune their practices through analyses of achievement data and problem solving. In what follows, we provide a detailed rationale for these two hypotheses.

According to Block and Pressley (2002), to comprehend written text a reader needs to be able to decode accurately and fluently and to have a wide and appropriate vocabulary, appropriate and expanding topic and world knowledge, active comprehension strategies, and active monitoring and fix-up strategies. So it follows that children who make relatively little progress may have difficulties in one or more of these areas. According to Slavin, Cheung, Groff, and Lake (2008), 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 that make expertise generalizable and enable students to self-regulate independently.

In addition, researchers have also identified the teacher’s role in incorporating cultural resources, including event knowledge (Ladson-Billings, 1994; Lee, 2007; McNaughton, 2002), and in building students’ sense of self-efficacy and more general engagement and motivation (Guthrie & Wigfield, 2000; Wang & Guthrie, 2004). Quantitative and qualitative aspects of teaching convey expectations about students’ ability that affect their levels of engagement and sense of being in control. Culturally and linguistically diverse students seem to be especially likely to encounter teaching that conveys low expectations (Dyson, 1999). A number of studies that focused not directly on reading comprehension but on schooling improvement have shown how these expectations can be changed and how they influence instruction and learning. In general, both changes to beliefs about students and more evidence-based decisions about instruction are implicated, often in the context of schoolwide or even clusterwide initiatives (Bishop, Berryman, Tiakiwai, & Richardson, 2003; Phillips et al., 2004; Taylor et al., 2005).

It follows that low progress could be associated with a variety of teaching and learning 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 may therefore have a context-specific profile. Although our research-based knowledge means 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, say in a cluster of similar schools serving similar communities, have difficulties associated with decoding or using comprehension strategies or both, and how the teaching that occurs in those schools is related to those difficulties. Buly
and Valencia (2002) provided a case study from a policy perspective on the importance of basing any intervention on specific profiles, rather than on assumptions about what children need (and what instruction should look like). In that study, a policy mandating phonics instruction for all students in the state of Washington who fell below literacy proficiency levels was shown to have missed the needs of the majority of students, whose decoding was strong but who struggled with comprehension or language requirements for the tests.

In the context examined in the current research, several explanations were possible for the low levels of reading comprehension. One was that children's comprehension levels were low because of low levels of accurate and fluent decoding (Tan & Nicholson, 1997). A second explanation was that the children may have learned a limited set of strategies; for example, they may have been able to recall well but were weaker in more complex strategies for drawing inferences, synthesizing, and evaluating, or the children may not have been taught well enough to control and regulate the use of strategies (Pressley, 2002). Other possible contributing reasons might have been more language-based—that children's vocabulary may have been insufficient for the texts used in classroom tasks (Biemiller, 1999) or that the children were less familiar with text genres. The patterns of Matthew effects may also have been present in classrooms, where culturally and linguistically diverse children receive more fragmented instruction that is overly 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 was also a set of possible hypotheses around whether the texts, instructional activities, and the pedagogy of the classroom enabled cultural and linguistic expertise to be incorporated into and built on in classrooms (Lee, 2000; McNaughton, 2002). But each of these needed to be checked against the localized patterns of instruction in the classrooms in order for the relationships to be tested.

Our argument is that using detailed evidence to affect instructional changes in a sustained way requires more than an intervention that prespecifies these changes. As Pressley (2006) recently noted, we need to know about being effective at the school level too. Collaborative analysis around evidence from teachers' own classrooms has been implicated as an important component of professional development aimed at improving teaching and student achievement. In their review of effective professional development, Hawley and Valli (1999) identified such collaborative analysis as a more effective form of professional development than traditional workshop models.

Further research evidence suggests that approaches in which professional development focuses on joint problem solving around agreed evidence result in improvements in student achievement, particularly reading comprehension (e.g., Cawelti & Protheroe, 2001; Taylor et al., 2005). Cawelti and Protheroe identified teacher analysis and use of achievement data as critical factors responsible for student reading achievement gains in six formerly underperforming districts with successful school improvement efforts.
There are several critical features of professional learning communities that can effectively analyze evidence to improve teaching practices and raise student achievement (Coburn, 2003; Robinson & Lai, 2006; Toole & Seashore-Louis, 2002). 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 the present case, that of comprehension), but it also entails detailed understanding of the nature of teaching and learning related to that domain (Timperley, Wilson, Barrar, & Fung, 2007), such as the research-based evidence on effective teaching of comprehension described earlier (e.g., Block & Pressley, 2002). However, being theoretically rich requires not just consideration of researchers’ theories but the engagement of teachers’ tacit ones (Robinson & Lai, 2006; Timperley et al., 2007). Engaging the teachers’ theories uncovers the reasons and conditions that have resulted in their current practices. This allows the reasons to be challenged if, for example, they are based on inaccurate assumptions and permits the necessary conditions to be taken into account to improve practice (see Robinson & Lai, 2006, for more details). A recent review of the literature indicates that this process is strongly linked to interventions that have improved achievement (Timperley et al., 2007).

Not only do teachers’ theories need be engaged alongside researchers’ theories, but any theory competition needs be resolved without a particular theory being privileged (Robinson & Lai, 2006). This process increases the validity of the emerging theories by allowing for disconfirming evidence from all parties to be treated and tested equally, rather than privileging researchers or teachers’ theories. Robinson and Lai provided the framework by which different theories can be examined using four standards of theory evaluation: 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 incorporate new goals, values, and contextual constraints).

This framework also means that a second feature of an effective learning community is that the goals and practices of a community are based on evidence. That evidence should draw on descriptions of children’s learning as well as descriptions of instruction and teaching practices. However, what is also crucial is the validity of the inferences drawn or claims made about that evidence, as meanings and implications of evidence are not self-evident (Coburn, Touré, & Yamashita, in press). Robinson and Lai (2006) suggested that all inferences be treated as competing theories and evaluated using the evaluation framework detailed previously.

So this requires a further feature, which is an analytic stance to the collection and use of evidence. A research framework needs to show whether and how planned interventions do influence teaching and learning, enabling the community to know how effective interventions are in meeting goals. Therefore, the research framework adopted by the community needs to be staged so that the effect of interventions can be determined.
A final feature of an effective professional learning community is that the researchers’ and teachers’ ideas and practices need to be culturally located. By this, we mean that the ideas and practices that are developed and tested should entail an understanding of children’s language and literacy practices, as these reflect children’s local and global cultural identities. It is important to know how these practices relate (or do not relate) to classroom practices (McNaughton, 2002; New London Group, 1996).

In summary, we claim that sustaining accelerating rates of achievement in reading comprehension for culturally and linguistically diverse students in the schools of poorer communities is dependent on two developments: the first is the development of professional learning communities focused on critically analyzing the effectiveness of instruction, and the second is the fine-tuning of instruction to better meet the learning needs of students in the communities. We report here on a three-phase schooling improvement study with these elements that was aimed at accelerating achievement. Elsewhere we have described in detail the processes occurring in each of the phases (Lai & McNaughton, 2008; Lai, McNaughton, MacDonald, & Farry, 2004; McNaughton, Lai, MacDonald, & Farry, 2004). Here our focus is on the outcomes of the intervention in terms of trajectories of change over time and distributions of achievement across groups using the concept of acceleration.

Method
The project was designed as a collaboration involving schools in a New Zealand Ministry of Education schooling improvement initiative, the initiative leaders in the schools, the Woolf Fisher Research Centre (The University of Auckland) and representatives of the New Zealand Ministry of Education. The Ministry of Education representatives were long-standing members of the professional learning community and were invited by the schools to participate. They had no control over the funding for the research, and their role was to learn from the emerging results to facilitate greater research–policy–practice links (Annan, 2007). Seven Decile 1 schools (i.e., schools with the highest proportion of students from the lowest socioeconomic communities) from urban centers in the South Auckland area took part. Two of these schools were contributing schools (Year 1–Year 6); three were full primary schools (Year 1–Year 8); one was an intermediate school (Year 7–Year 8); and one was a middle school (Year 7–Year 9). The schools’ sizes ranged from 292 students to 593 students.

The schooling improvement initiative had been running for five years prior to this project. In those five years, the Ministry of Education and researchers have been working with schools on addressing student behavior and staffing, developing a shared focus on student outcomes, forming partnerships with each other and the community, and establishing more effective teaching and management practices (Annan, 2007; Lai, 2003).
Participants (Students and Teachers)

We report two overlapping groups of students. One group comprised all the students present at the beginning of the three-year study (baseline sample, \( N = 1,975 \) students). A second group comprised three cohorts of students, initially from School Year 4 (comparable to grade 3 in the United States; students were 9-year-olds), Year 5 (10-year-olds), and Year 6 (11-year-olds) who were followed longitudinally for three years (total \( N = 238 \)).

The baseline data (February 2003) were collected from the 1,975 students in six of the schools (one school that joined the project was unable to participate in the first round of data collection). The total group consisted of equal proportions of males and females (50% and 50%, respectively) from 14 ethnic groups. Four main groups from indigenous and ethnic minority Pasifika communities made up 87% of the sample. These groups were Samoan (33%), Māori (20%), Tongan (19%), and Cook Island (15%) ethnic groups. Approximately half the children had a home language other than English. A number of students dropped out and reentered the project during the period of the intervention (\( n = 536 \)), with the majority of those students having missed at least half of the intervention.

The second group comprised three cohorts of students who were followed longitudinally from Time 1 to Time 6 and who were present at all six data collection time points. The total of 238 students is indicative of the high turnover of students in these schools, between 25% and 30% on average per year. The three cohorts were Cohort 1 (\( n = 114 \)) students who were Year 4 at Time 1; Cohort 2 (\( n = 56 \)) students who were Year 5 at Time 1, and Cohort 3 (\( n = 68 \)) students who were Year 6 at Time 1. These students were a subset of the students included in the baseline sample.

Approximately 70 teachers were involved in each year of the project, including school-appointed literacy leaders who were normally the deputy or associate principal. Characteristics of the teachers varied somewhat year to year, but in general around two thirds had five or more years of experience, and 10% were beginning teachers. Eleven percent taught in bilingual classes (including Samoan, Tongan, and Māori bilingual classes). A third of the teachers were Pasifika (from different Pacific Island communities) or Māori (indigenous).

Design

At the core of the following analyses is a quasi-experimental design from which qualified judgments about possible causal relationships are made. 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. This variability is inherent to human behavior generally (see Sidman, 1960) and specifically is present in applied settings (Risley & Wolf, 1973). These circumstances require a design that deliberately incorporates variability and the sources of the variability and that has longitudinal properties. The quasi-experimental design described is appropriate to the circumstances of testing effectiveness over a period of time, given that variability is an important property (Raudenbush, 2005).
It was appropriate also given that the question is the effectiveness of the intervention in terms of an expected rate of gain for a particular cluster of schools if the intervention had not occurred and given that the obtained rate of gain should match a national distribution.

Repeated measures of children’s achievement were collected in February 2003 (Time 1), November 2003 (Time 2), February 2004 (Time 3), November 2004 (Time 4), February 2005 (Time 5), and November 2005 (Time 6) as part of the quasi-experimental design. 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 4–5–6–7–8), which provides a baseline forecast of what the expected trajectory of development would be if the planned interventions had not occurred (Risley & Wolf, 1973). Successive stages of the intervention can then be compared with the baseline forecast and judgments about acceleration that are contextually valid can be made. In the present case, the first of these planned interventions was the analysis and discussion of data. The second was the development of instructional practices through workshops. 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 and the external validity comes from the systematic analysis of replication across cohorts within the cluster.

Two sorts of general analyses are possible with the repeated measures. Analyses can be conducted within each year. These are essentially pre- to post-test measures. However, because the measures we use can 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 and hence a measure of acceleration. A more robust analysis of relationships between the intervention and achievement is provided by repeated measures within the quasi-experimental design format. They show change over repeated intervals and acceleration relative to expected change across school months and summer breaks.

Good science requires replications (Sidman, 1960). In quasi-experimental research, the need to systematically replicate effects and processes is heightened because of the reduced experimental control gained with the design. This need is specifically identified in discussions about alternatives to experimental randomized designs (Borko, 2004; Chatterji, 2005; Raudenbush, 2005). For example, McCall and Green (2004) argued that in applied developmental contexts, evaluation of program effects requires a variety of designs including quasi-experimental, but our knowledge is dependent on systematic analyses across sites. Replication across sites can add to our evaluation of program effects, particularly when it is inappropriate or premature to conduct experimental randomized designs. Such systematic replication is also needed to determine issues of sustainability (Coburn, 2003) and scaling up (McDonald, Keesler, Kauffman, & Schneider, 2006).
The design used with this cluster of schools has in-built replications across age levels within the quasi-experimental design format. These provide a series of demonstrations of possible causal relationships. However, there are possible competing explanations for the conclusions of the clusterwide results that are difficult to counter with the quasi-experimental design. These are the well-known threats to internal validity, two of which are particularly threatening in the design adopted here. (Some of the 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, and we present analyses of these subgroups; similarly repeated testing occurred but with instruments that were designed for the interval of repetition and with alternative forms.)

The first of these major threats is that an unknown combination of factors unique to these schools and this cluster—that is, the immediate historical, cultural, and social context for these schools and this particular cluster—determined the outcomes; technically, this is partly an issue of ambiguous temporal precedence and partly an issue of history and maturation effects (Shadish, Cook, & Campbell, 2002). For example, the nature of students might have changed in ways that were not captured by the general descriptions of families and students, or perhaps, given that the immediate history included a number of New Zealand Ministry of Education initiatives (Annan, 2007), the schools were developing more effective ways of teaching anyway.

A second major threat is that the students who were followed longitudinally and were continuously present over several data points were different in achievement terms from those students who were only present in the baseline and subsequently left. It might be, for example, that the comparison groups contained 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 and Isken (1996) reported 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 (Shadish et al., 2002). As the projections on the projected baseline are 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 any real effects.

There are three ways of adding to the robustness of the design in addition to the in-built replications that meet the two major threats. The first is to use, as a comparison, a similar cluster of schools that has not received the intervention. It was possible to identify such a cluster post hoc and examine the baseline levels in schools after a year of intervention against the levels in the second cluster that had not experienced the intervention. The second cluster of schools was in a neighboring suburb that was part of a different Ministry of Education schooling improvement initiative. The second cluster was similar in geographical location, in type (all Decile 1 schools), in number of schools ($n = 7$), in number of students ($n = 1,161$), in ethnic and gender mix (equal proportions of males and females...
from more than 12 ethnic groups; the major groups being Samoan [37%], Māori [22%], Cook Island [18%], and Tongan [15%]), in starting levels of achievement, and in prior history of interventions. This cluster was to participate in the same three-year intervention as part of a staggered replication series and was scheduled to begin the intervention a year later than the one reported here. Therefore, all target year levels (School Year 4–School Year 8) in the second cluster were measured exactly one year after the baseline was established in the first cluster reported here (Lai, McNaughton, MacDonald, Amituanai-Toloa, & Farry, 2006).

The two achievement baselines are shown in Figure 1 and Figure 2. The baselines are cross-sectional data collected at the beginning of the first year of the intervention in each cluster but a year apart. They are in the form of normalized scores (stanines) from the Supplementary Tests of Achievement in Reading (STAR; Elley, 2001). The raw scores have been converted into 9-point stanine scores based on school level norms. The comparison shows that a year after the intervention had started in the first cluster of schools, the second cluster of schools continued to have levels of achievement similar to those that the first cluster of schools originally had.

This comparison adds to the design conclusions by establishing that there was not some general impact on similar neighboring area Decile 1 schools operating over the time period of the intervention and that if no intervention had taken place in the first cluster of schools, the levels of achievement would have remained around stanine 3.0.

A second way of adding to the believability 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

Figure 1. Baseline for Cluster 1 (2003): Student Achievement Across Year Levels (n = 1,975)
in subsequent repeated measures. We accomplished this for the first year data by checking the achievement data for those students who were present at two time points (Time 1 and Time 2) versus those students who were only present in the cross-sectional baseline established at Time 1 (Time 1 only).

The results of this checking are given in Table 1 for raw scores. The comparisons indicate that in each of all but one comparison, the two groups of students were not significantly different.

These two additional checks add to the robustness of the design by showing 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

Table 1. Reading Achievement (Raw Scores on STAR): Students Present Only at Time 1 (February 2003) and Present at Both Time 1 and Time 2 (November 2003) by Year Level

<table>
<thead>
<tr>
<th>Raw Score</th>
<th>Time 1 Only</th>
<th>Time 1 and Time 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>M</td>
</tr>
<tr>
<td>Year 4</td>
<td>34</td>
<td>16.26</td>
</tr>
<tr>
<td>Year 5</td>
<td>34</td>
<td>18.94</td>
</tr>
<tr>
<td>Year 6</td>
<td>30</td>
<td>23.6</td>
</tr>
<tr>
<td>Year 7</td>
<td>33</td>
<td>30.61</td>
</tr>
<tr>
<td>Year 8</td>
<td>34</td>
<td>32.68</td>
</tr>
</tbody>
</table>

* p < 0.05.
of interventions and resourcing. In addition, they do not support the competing explanation that the students analyzed in the longitudinal design were higher achievers anyway and hence any “progress” simply reflected their usual levels of reading achievement compared with all other students.

The design had three phases, each occurring over the course of a year. The first phase was the collaborative analysis of evidence phase, the second phase added to this targeted workshops, and in the third phase, a program of building the professional learning community replaced the workshops while the critical analysis introduced in the prior two phases continued. In repeated-measures design terms, these could be considered an A - A + B - A + C design.

Procedures

Phase One: Analysis of Data, Feedback, and Critical Discussion. The first phase introduced both hypothesized components (analysis and change in instructional practices) within the professional learning community without targeted workshops. Areawide data on both achievement and instruction were first analyzed by the school leaders and researchers in two meetings and then analyzed by senior managers and senior teachers with each school using their specific school data. Additional meetings at school level were conducted with support from Lai.

The format of the meetings was identical: The researchers presented achievement and teacher observation data collected as part of the intervention and then facilitated a discussion about the data and their implications for classroom practice. The analysis, feedback, and discussion process involved two key steps. The first step was a close examination of students’ strengths and weaknesses and of current instruction to understand learning and teaching needs, and the second was a discussion of competing theories about the problem and evaluation of the evidence for these competing theories. This meant using evaluation standards of accuracy, effectiveness, coherence, and improvability (Robinson & Lai, 2006). This process ensured that the collaboration was a critical examination of practice and that valid inferences were drawn from the information.

An example of a data discussion using this process is as follows: In New Zealand, there has been considerable debate about the causes of low reading achievement. One school of thought is that it is primarily constrained skills, such as students’ decoding, that are the cause of the low reading achievement whereas another view is that it is primarily unconstrained skills that are the cause. In a meeting, the professional learning community tested these two theoretical positions (the latter being held predominantly by the school leaders in the community) by carefully examining profiles of students’ needs from the achievement data. In other words, the community engaged and tested teachers’ and researchers’ theories using the standard of accuracy (empirical claims about practice are well founded in evidence). The profiles indicated that students were high decoders but were weak in other aspects of reading comprehension, thereby ruling out one of the opposing theories (decoding was the reason for low comprehension) and ruling in the other (students could decode but not comprehend texts). In addition,
analysis of achievement and classroom observation data revealed other more pressing issues. Students were not checking the text for evidence when comprehending, and teachers were not prompting students to do so (Lai & McNaughton, 2008). So the community realized that they did not need an intensive focus on decoding but rather a focus on the other areas of reading comprehension, such as checking for textual evidence. The feedback procedures with examples are described fully in Lai and McNaughton (2008) and Robinson and Lai (2006).

Prior to the project, a review of each of the school leaders’ capacity to critically analyze student achievement data was undertaken as part of the New Zealand Ministry of Education’s assessment of the initiative (Lai, 2003). In that review, school leaders were evaluated twice. They were first required to bring student achievement data from their school and demonstrate to the reviewer in an interview that they could analyze it accurately and make appropriate links to aspects of their teaching practice. They were then required to write a case study documenting how their school had used student achievement data to change an aspect of school practice to raise achievement. The results of those two evaluations indicated that the school leaders were accurate in their analysis of achievement data but needed further support in making links based on evidence between their achievement information and their practices. The first phase of this project therefore built on these existing skills and extended them further by helping teachers to make connections between specific teaching practices and student achievement.

**Phase Two: Targeted Professional Development.** The collection, feedback, and discussion of profiles continued in the second phase in the second year, but in addition targeted professional development consisting of 10 sessions over half a year occurred. The professional development was designed from the profiles and known dimensions of effective teaching and was led by McNaughton. The curriculum for the sessions was a mixture of theoretical and research-based ideas as well as teachers’ investigations and examples from their own classrooms.

Session 1 introduced theoretical concepts of comprehension and related these to the profiles of teaching and learning. A theoretical model was presented that drew on Sweet and Snow (2003) and developmental analyses such as Whitehurst and Lonigan (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 2 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. Session 3 introduced theories and research relating the role of vocabulary in comprehension. Readings used were Biemiller (1999), Pressley (2000), and others that identified features of effective teaching of vocabulary. The task for this session was to design a simple study carried out in the classroom that built vocabulary through teaching. Sessions 4 and 5 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 on the range and types of books available in classrooms and engagement by different students. Sessions 6 and 7 introduced concepts of incorporating 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 analyzing these features of instructions were set. In Sessions 8 and 9, transcripts of the video classroom lessons were used to exemplify patterns of effective teaching in different settings, such as guided reading and shared reading, and the practice of examining and critiquing each others’ practices was developed. The ninth session also covered some specific topics that the groups had requested, such as the role of homework and teaching and learning in bilingual settings. Session 9 also involved planning to create learning circles within schools, where colleagues observed in each others’ 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 added to the two core components in several ways. The cluster collection and feedback and critical discussion of achievement data continued. In addition, the school leaders continued to guide the learning circles developed in the professional development phase, focusing on the dimensions of instruction developed through the sessions. A major new feature was the development and use of planned inductions into the focus and patterns of teaching and professional learning in the schools. The schools experienced staff turnover of differing degrees from year to year, 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.

Another new feature was a teacher-led conference, designed to build the effectiveness of the professional communities across schools even further. School teams developed action research projects often with pre- and posttesting components to check various aspects of their programs. The questions for these projects were generated by teams within schools. The researchers helped shape the questions and the processes for answering the questions. Two research meetings took place at each of six schools (the seventh had a change of principal and literacy leader and declined to develop projects, although staff attended the conference). Several of the research topics were concerned with increasing vocabulary both in language programs and in instructional reading and writing programs. Others included increasing factual information in narrative writing (to build awareness of use of factual information), teaching of skimming and scanning in the reading program, use of instructional strategies to increase the use of complex vocabulary in writing, the effects of using a new assessment tool for writing to inform
teaching, redesigning homework to raise literacy levels, and the use of critical thinking programs. In each case, the projects involved use of formal or informal assessments of student outcomes. A total of 11 projects were presented (conference format) at a teacher-led conference in the fourth term of the school year that 90% of the teachers attended. Other professional colleagues such as literacy advisors attended the conference also.

**Measures**

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 STAR (Elley, 2001). These tests were designed for repeated measurement, are used by schools, and provide a recognized, standardized measure of reading comprehension that can be reliably compared across schools. Both tests have high reliability and validity (Elley, 2001; Reid & Elley, 1991). In addition to these assessments, the schools used other reading measures for both diagnostic and summative purposes, and the baseline results for these are reported elsewhere (McNaughton et al., 2004).

The revised PAT in Reading measures both factual and inferential comprehension of prose material in Years 4–9. Each prose passage consists of 100–300 words and is followed by four or five response options that are multichoice. The prose passages are narrative, expository, and descriptive, and different year levels complete different combinations of prose passages. Different year levels complete different parts of the PAT. The proportion of factual to inferential items per passage is approximately 50%–50% in each year level. Depending on test parts, the test–retest reliability ranged from 0.85 to 0.88, and the split-half reliability ranged from 0.88 to 0.92.

The STAR was designed to supplement the assessments that teachers make about students’ close reading ability in Years 4–9 in New Zealand (Elley, 2001). It has parallel forms and can be given at three points during the school year. The raw scores can be converted to 9-point stanine scores based on New Zealand national norms. In Years 4–6, the test consists of four subtests measuring word recognition (decoding of familiar words through identifying a word from a set of words that describe a familiar picture), sentence comprehension (completing sentences by selecting appropriate words), paragraph comprehension (replacing words that have been deleted from the text in a Cloze format), and vocabulary range (finding a synonym for an underlined word). Only the paragraph-comprehension subtest is not multichoice and consists of 20 items, 10 more than in the rest of the subtests. In Years 7 and 8, students complete two more subtests, which involve understanding the language of advertising (identifying emotive words from a series of sentences) and reading different genres or styles of writing (selecting phrases in paragraphs of different genres that best fit the purpose and style of the writer). In Years 7 and 8, there are 12 items per subtest except for in the paragraph-comprehension subtest, which consists of 20 items. The test–retest reliability was 0.91, and the split-half reliability was 0.91 for the total scores.
The correlation between the two tests was 0.62 ($p < 0.01$). Elley (2001) reported correlations between 0.70 and 0.78 for Years 4–8 students on these tests, indicating that the tests measure similar but not identical facets of reading comprehension. After the initial testing, schools focused on using the STAR data, and the outcome data on reading comprehension reported here for the overall project are from across the six time points in which the STAR data were used.

**Reliability**

At the beginning of the project, the schools and researchers developed an intra-school standardized process of administering the test and moderating the accuracy of teacher scoring. This involved standardizing 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 the Woolf Fisher Research Centre during data entry and during analysis. The STAR and PAT were administered as part of schools’ normal assessment cycle at the beginning of the school year, and thereafter STAR was administered at the end of each year also (using the parallel form). Additional assessments conducted at Time 1 (February 2003) involved analyzing student scores on factual and inferential questions from the PAT and from the STAR and qualitatively coding the types of errors that students made on the Cloze passage according to the types of errors reported in the STAR manual (Elley, 2001). Four raters were trained to code errors. These raters subsequently discussed how to code the errors and collectively rated a sample of tests so that the reliability of coding could be determined. The coding was subsequently checked, and interobserver agreement on 10% of students’ subtests (across ages) was 90.5% (for more details, see Lai et al., 2004).

**Description of the Instructional Program**

Observations of reading sessions in a sample of classrooms were conducted early in the first year to provide a general description of classrooms and a sample of how features of teaching and learning might map onto the achievement data. The initial observations involved in situ running records of each component of the reading lessons by trained observations and tape recordings of selected group work in 15 classrooms. Subsequent observations in the second year involved systematic observations from video transcripts of classroom lessons using a coding protocol based on the focus of the instructional program (see Lai et al., 2004; McNaughton, Amituanai-Toloa, & Lai, 2008).

Generally, the program was similar across classes and schools and similar to the general descriptions of the New Zealand teaching in the middle grades (Ministry of Education, 2006; Smith & Elley, 1994). Class sizes generally ranged from 21 to 26 students. A 10- to 15-minute whole-class activity, which involved mostly introducing and sharing a text, often a narrative text, or reviewing the previous day’s work, was then usually followed by a 30- to 40-minute guided reading session in small groups led by the teacher using an instructional text.
These included text study and analysis (such as studying plot or character in narrative texts and extracting and using information in informational texts), specific group or paired forms of instructional/guided reading (such as reciprocal teaching, Palincsar & Brown, 1984), and individual or group project work (such as developing taxonomies of categories introduced in science topics). Typically, the teacher worked with two groups over this time period and held conferences on the run with other groups.

Levels of engagement were generally high, with routines well established and many instances of teacher–student and student–student interactions. The general organization meant whole-class activities occurred three to five days per week and small-group work with one or two groups occurred often daily, so that each group had at least one session but up to three sessions with direct teacher guidance each week. However, there was some variation in frequency of contact with each group among schools. When not with the teacher, groups did a range of activities. Some had developed to the point of being able to operate just with peer guidance in reciprocal teaching arrangements (Palincsar & Brown, 1984). In most classrooms, worksheets sometimes related to texts were used, which contained questions about a text and often contained sentence, word, or subword studies.

The detailed observations were used together with the assessments and achievement data to generate hypotheses about instructional changes. The hypotheses followed partly from a general view that a new program of instruction was not needed, rather existing practices needed to be fine-tuned. The reasons for this view included evidence from the baseline that instruction was sufficiently effective to enable students to make about a year’s gain for one year at school across the age levels, but on average students consistently remained two years below at each level. Also, the national and international data relating to literacy instruction and achievement in New Zealand suggest generally effective instruction that is not as effective with Māori and Pasifika students (Alton-Lee, 2004).

One hypothesis was that more explicit and incidental instruction of vocabulary was needed. This was not surprising, given that about half of the students had a home language other than English. The evidence from the profiling tests was that the students had a limited vocabulary range, and vocabulary instruction was seen as essential to boosting reading comprehension (Pearson, Hiebert, & Kamil, 2007). However, classroom observations revealed high rates of explicit instruction. Analyses suggested that these rates reflected a focus on technical and topic-related words and that a focus was needed on more literary or academic words of idiomatic usage, figurative language, and familiar words used in unfamiliar ways (Pearson et al., 2007).

A second hypothesis was about strategy instruction. Analyses of achievement tests revealed difficulties with comprehending connected text. A solid research base provides considerable evidence for the significance of developing strategies (Pressley, 2002). However, the analyses of classroom instruction revealed the presence of high rates of explicit teaching of strategies. More analyses indicated a specific problem with strategy instruction: the limited use of text evidence to
detect confusions or threats to meaning or to check and corroborate meanings. The hypothesis was that strategies such as prediction and clarification needed to be linked to checking and discussing meanings from texts.

A third hypothesis had two components. One was that students’ learning would be improved if instruction capitalized more on students’ cultural and linguistic resources (Ladson-Billings, 1994; Lee, 2007). That is, more effective instruction would incorporate children’s event knowledge and interests through judicious selection and matching of texts and use of familiar language and culturally based forms of teaching. The evidence was that this did occur in classrooms but that a more deliberate focus was needed. However, complementing this form of teaching would be instruction that increased students’ awareness of the goals and formats of classroom activities and the relevance of their skills and knowledge to these activities. Classroom observations had shown instances in which students were confused about what they were required to do.

The fourth hypothesis concerned the general issue of practice (Sweet & Snow, 2003). Each of the previous hypotheses is dependent on the extent of practice within texts and planned variation in exposure across types of texts. Observations also revealed a tendency for a large proportion of time to be spent in explicit instruction outside of actual reading and for the teacher to be dominant in extended interactions of questioning, often involving Initiation Response and Evaluation sequences (Cazden, 2001), which took time away from reading. Adolescent readers identify independent reading and access to high-quality and diverse reading materials as especially motivating (Ivey & Broaddus, 2001).

These hypotheses and the observations on which they were based were systematically discussed along with the achievement data with the teacher leaders and the staff in the schools in the first year and formed the basis for specific professional development in the second year. Further observations conducted at the beginning and end of the second year provided a check on the integrity of the intervention at the level of classroom instruction. A full description is contained in McNaughton et al. (2008). The observations showed that changes had occurred in most areas consistent with the hypotheses: an increased focus on vocabulary, increased use of strategies to maintain and check meanings, increased instruction to cue students’ awareness of tasks and strategies, and increased presence of the four elements of the instructional focus in each teacher’s instructional exchange. The planned focus on incorporating students’ background had not increased beyond levels present at the beginning of Year 1.

**Data Analysis**

Repeated measures using both raw scores and normalized scores in stanines provide descriptions of achievement patterns over time. Standard statistical tests for mean differences were used as appropriate, such as Hotelling’s $T^2$ tests (corrected through Bonferroni procedures as needed), chi-square tests, effect sizes (Cohen’s $d$), and multivariate analyses of variance. The calculation of the effect size indexes (Cohen’s $d$) was based on Cohen’s 1988 and 1992 equations:
National expectations were determined from the national norms provided in the assessment tool and the test developer’s guidelines for expected national progress. The stanine scale is age adjusted, so it would be expected that students making expected progress over time would remain at the same stanine level. Data from all testing periods were recorded and stored in SPSS, which was used as the primary analysis tool for these statistics.

Testing the effectiveness of the interventions in terms of acceleration proceeded in several steps. The first step was to establish the effectiveness within the limits of the quasi-experimental design. We use the logic of the quasi-experimental design described earlier and compare actual levels of cohorts and the total group against what was projected by the baseline. As described earlier, we have increased the robustness of this basic design in several ways, including replications across cohorts and checks on subject bias and on history and maturation effects via a lagged comparison with a cluster of similar schools. In addition, we examine the distribution of achievement for the cohorts of students at Time 1 and their distributions after three years at Time 6. This is essentially a pre- to posttest analysis, but because we have used stanines that are normalized and therefore age adjusted, any consistent shifts mean that change has occurred relative to normal expected growth.

Testing the generalized effectiveness across subgroups in the cohorts followed longitudinally involved comparisons between subgroups. We first examined differences across ethnic groups. The comparison based on ethnic groups was binary. Students were categorized into two ethnic groups: Māori for indigenous New Zealand Māori (20% of the sample) and non-Māori, who were primarily students from Pacific Island nations (74% of the sample). There were insufficient numbers of students from the other ethnicities to allow comparisons with other ethnic groups (New Zealand European, 2%; Asian, 3%; and other, 1%). The analyses of achievement group differences were also binary. Low-start students were those in the “low” and “below average” bands of achievement in the baseline test as indicated by the STAR manual (Elley, 2001). High-start students were those in the “average,” “above average,” and “excellent” bands of achievement as indicated by the STAR manual. We were unable to analyze the data according to low, medium, and high start due to the few number of students (< 10) in the high-start bands at the beginning of the study.

Modeling the Data. We attempted to fit growth curves to the longitudinal data (over six data points) to model the changes and to analyze subgroups. There were six repeated measures for each student with many missing values, resulting in a total of 6,117 records in the data set across seven schools. The obvious hierarchical structure of such data would be students (Level 1) nested in classrooms (Level 2) nested in schools (Level 3). Two difficulties with this structure stem
from the characteristics of classes. Class 1 in Calendar Year 1 was not the same class as Class 1 in Calendar Year 2 because it often would have a completely different set of students and possibly a different teacher. We could impose a hierarchical structure by replacing the class label by the class–calendar year label. But even then, there are difficulties brought on by the fact that some students changed classes between the first and the second tests in a given calendar year. The indications are that 25 students changed classes in Calendar Year 1, 129 changed classes in Calendar Year 2, and 283 changed classes in Calendar Year 3. We could remove these students, but these changes are not anomalous; rather they are a typical feature of the schools, and moreover, this would result in the loss of at most 437 students (out of 6,117 records of data)—“at most” because there may be overlap among the groups changing classes in the three calendar years. Dropping these students may cause a biasing effect on the data.

As a consequence, we attempted to develop an appropriate model, taking account of as much of the hierarchical structure as is actually present, leaving both school and class in the first model. We found no variance attributable to school and almost no proportion of variance attributable to classes (< 10%). Given the issues with defining the class label, and the fact that so little of the variance was attributable to school and class, we dropped those from subsequent models of growth over the six data points.

The data set we used for the subsequent models were a subset of the total data set, composed of only the students with all six data points. The students in this subset were not statistically significantly different from other students at the baseline (see previous section on Design). However, from Year 2 onwards, there were statistically significant differences between these students and a number of students with missing values who were either absent or transient (Lai, McNaughton, & Timperley, 2007). This is most likely attributable to the effects of the intervention, given the lack of statistically significant differences at baseline. Hence, our model only included students with all six time points to avoid any confounding effects from students with differential exposure to the intervention. We also did not include in our modeling two schools (intermediate and middle school) that would have, at most, cohorts participating in two years of our intervention.

It would have been possible to use imputation methods, such as multiple imputation, to handle the missing data, but the assumption of missing-at-random missingness and monotone missingness (Horton & Kleinman, 2007) was not applicable to the data collected. The purpose of the modeling was to describe the growth in reading skills over the three-year period for the intervention, so including students’ data that were not missing at random could bias the model. So we decided that the analysis would lean on the conservative side and base the growth model on those students who had completed all six tests to avoid the potential bias if we misspecified the imputation model.

From this data set, we first attempted to model growth over the six data points (three years) using the statistical package Minitab. We began by testing a
model that assumed no theoretical reason for a difference in the interval of testing between school years, in this case between Test 2 (end of year) and Test 3 (beginning of year), between Test 4 (end of year) and Test 5 (beginning of year), and within school years. We tested for goodness of fit for a linear model, treating time as a continuous covariate against the six-parameter model. Given the test statistic $F(4, 1149) = 6.89, p < 0.001$, we rejected this model. Similarly, we tested a quadratic (in time) model against the six-parameter model and obtained $F(3, 1149) = 8.38, p < 0.001$. Hence, the quadratic model was also rejected.

In the research literature, one model of growth over time suggests a staircase-shaped model, with achievement increasing in the calendar year and decreasing or increasing (depending on the population being examined) over the summer holidays (Borman, 2000). With students similar to ours, Borman found a decrease in achievement over summer. However, a preliminary analysis of our data for the cluster of schools suggested that achievement reached a plateau instead, with no further gains or losses over the first summer (Lai & McNaughton, 2008). So we tested a subsequent model that took into account the possible differences between and across the school year. With the subsequent, generalized model, we could predict that given our intervention, we would improve achievement across the school year, but that average achievement would plateau (students would not make any progress) between school years.

To gain a more detailed understanding of our data and to test whether the intervention process successfully avoided having differential effects, we applied hierarchical linear modeling (HLM) with repeated measures, as implemented by the MIXED procedure in SAS, as an extension to the generalized model. The variances and correlation structure of the repeated measures were modeled first and then the generalized least-square estimates of the factors, such as students’ initial achievements and cohort groups; time differences were obtained by optimizing residual likelihood functions with the Newton–Raphson algorithm (Littell, Henry, & Ammerman, 1998; Wolfinger, Tobias, & Sall, 1994) and Satterthwaite approximation for the modifications of the degrees of freedom. Throughout the process of model extensions, the Akaike’s information criterion (AIC) and Bayesian information criterion (BIC) were used to determine which combination of factors provided better models.

## Results

**Establishing the Baseline and Initial Profiles**

At the beginning of the project, the stanine distributions of both tests, STAR and PAT, indicated that the average student experienced considerable difficulty on these measures of reading comprehension (see Figure 3). The average student in both tests scored in the below-average (Stanines 2 and 3) band of achievement. For both the PAT and STAR tests, the mean stanine was 3.10, indicating that achievement was about two years below-average levels. Over 60% of students scored in the low (Stanine 1) or below-average (Stanines 2 and 3) bands,
and less than 5% were in the above-average (Stanines 7 and 8) or superior bands (Stanine 9). Twenty-three percent of students would be expected at these lower (Stanines 1–3) or higher (Stanines 7–9) extremes.

At Time 1 across the year levels, the pattern was the same in both tests, with the median in every year level at Stanine 3. Figure 1 (see page 37) shows these cross-sectional data, and the means and standard deviations for these cross-sectional data are given in Table 1. The near stable pattern across year levels indicates that the students made on average one year of progress for one chronological year at school (including both school months and summer months). Two implications can be drawn from this: The teaching was sufficient to maintain expected progress, albeit consistently at two years below average levels, and the teaching was not effective enough to accelerate progress.

**Design-Based Analysis of Acceleration**

Gains were analyzed, first as parallel comparisons with the projected means for each year established by the cross-sectional baseline, and then against a comparison group of similar schools. The three-year longitudinal data for cohorts plotted against the initial baseline are shown in Figure 4. Visual inspection suggests that the cohorts accelerated gains against the projections provided by the

---

**Figure 3. Progressive Achievement Test (PAT) and Supplementary Tests of Achievement (STAR) in Reading Stanine Distribution Across All Year Levels at Time 1 (February 2003)**
cross-sectional data after one year, two years, and three years. Planned comparisons after one year and after two years confirm the visual inspection.

After one year of the intervention, all cohorts were statistically significantly higher than the baseline projected from the cross-sectional data at Time 1 (see Table 2). This provides the initial design-based evidence that the intervention can be systematically attributed to the intervention.

Further analysis shows that after two years of the intervention, all cohorts were not just statistically significantly higher than the projected baseline (see Table 3) but that the difference between their scores and those of students in the same year level two years previously had increased. The effect sizes were now up to 0.59 as compared with 0.44 after one year of the intervention. This suggests that the intervention had a cumulative and positive effect on achievement.

As noted in the description of the design, we added two features to increase the robustness of the design. The first was to test the issue of subject selection bias, and we showed that the students in the longitudinal cohorts did not generally differ from all students in terms of initial achievement levels (see Table 1).
A second design feature was to compare the baseline projections with a cross-sectional baseline from a similar cluster of schools after a year had elapsed, thereby controlling for general history and maturation and other associated confounding variables. We have compared the outcomes of the intervention in the first cluster with the baseline of the second cluster established at the same time. After one year of the intervention, all year-level cohorts scored statistically significantly higher than the comparison cluster that had not experienced this intervention (see Table 4). The effect sizes were between 0.33 and 0.45. After two years,
all year-level cohorts continued to score statistically significantly higher than the comparison cluster that had not experienced this intervention, with effect sizes between 0.41 and 0.61 (see Table 5).

**Distributional Analysis of Acceleration**

The criterion for acceleration included shifting the distribution to match national norms. The following analyses examine the shift in distribution from Time 1 to Time 6 against the expected distribution.

### Table 4. Mean Student Achievement in Comprehension (in Stanines) After One Year of Intervention Compared With Cross-Sectional Comparison Cluster

<table>
<thead>
<tr>
<th>Comparison Year</th>
<th>Cross-Sectional Baseline Cluster 2 (February 2004)</th>
<th>Intervention Cohorts After One Year of Intervention (Time 3, February 2004)</th>
<th>t</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 5</td>
<td>M 3.39 SD 1.53 N 248</td>
<td>Year 4→5 M 3.96 SD 1.35 N 114</td>
<td>3.42**</td>
<td>0.40</td>
</tr>
<tr>
<td>Year 6</td>
<td>M 3.32 SD 1.51 N 237</td>
<td>Year 5→6 M 3.84 SD 1.62 N 56</td>
<td>2.28*</td>
<td>0.33</td>
</tr>
<tr>
<td>Year 7</td>
<td>M 2.71 SD 1.22 N 360</td>
<td>Year 6→7 M 3.26 SD 1.22 N 68</td>
<td>3.43**</td>
<td>0.45</td>
</tr>
</tbody>
</table>

* p < 0.05. ** p < 0.01.

### Table 5. Mean Student Achievement in Comprehension (in Stanines) After Two Years of Intervention Compared With Cross-Sectional Comparison Cluster

<table>
<thead>
<tr>
<th>Comparison Year</th>
<th>Cross-Sectional Baseline Cluster 2 (February 2004)</th>
<th>Intervention Cohorts After Two Years of Intervention (Time 5, February 2005)</th>
<th>t</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 6</td>
<td>M 3.32 SD 1.51 N 237</td>
<td>Year 4→5 M 4.04 SD 1.45 N 114</td>
<td>4.21***</td>
<td>0.49</td>
</tr>
<tr>
<td>Year 7</td>
<td>M 2.71 SD 1.22 N 360</td>
<td>Year 5→6 M 3.23 SD 1.29 N 56</td>
<td>2.95**</td>
<td>0.41</td>
</tr>
<tr>
<td>Year 8</td>
<td>M 2.75 SD 1.26 N 305</td>
<td>Year 6→7 M 3.53 SD 1.29 N 68</td>
<td>4.6***</td>
<td>0.61</td>
</tr>
</tbody>
</table>

** p < 0.01. *** p < 0.001.
Gains Time 1–Time 6, Total Group and Year Cohorts. There was a statistically significant acceleration in achievement for the total cohort ($n = 238$) from Time 1 to Time 6 of 0.97 stanine. The total gain represents about one year’s progress in addition to expected national progress over the three-year period (Table 6 presents the mean stanine and raw scores for the cohorts of students tracked across the three phases of the project, and this is shown graphically in the form of an overall shift in the stanine distribution in Figure 5). Each cohort also made significant gains compared with expected gains. By the end of the project, the average student scored in the average band of achievement, and now 10% of the students were in the above-average and superior stanine bands (Stanines 7–9). All cohorts made statistically significant accelerations in achievement across the three years, with the effect sizes for the age-adjusted scores between 0.36 and 0.76.

The obtained stanine distributions at Time 1 and Time 6 can be compared with the expected (normal) distribution. Because of small cell sizes at Time 1 in the banded stanines, the band cells were combined into two cells: number of students in Stanines 1–3 and number in Stanines 4–9. The distributions at Time 1 were significantly different from the expected norms, $X^2(1, N = 238) = 176.5, p < 0.001$. At Time 6, the obtained distribution yielded $X^2(1, N = 238) = 9.72, p < 0.01$. When a much more stringent test is used rather than the conventional $p < 0.05$ criteria, the observed distribution at Time 6 was not significantly different from the expected. However, the difference detected with the conventional, less stringent test ($p < 0.05$) supports the conclusion that more gains are needed to better approximate the expected distribution.

Table 6. Stanine and Raw Score Means by Cohort at Time 1 (February 2003) and Time 6 (November 2005)

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Stanine</th>
<th>Raw Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time 1</td>
<td>Time 6</td>
</tr>
<tr>
<td>Cohort 1</td>
<td>M</td>
<td>3.41</td>
</tr>
<tr>
<td>(N = 114)</td>
<td>SD</td>
<td>1.32</td>
</tr>
<tr>
<td>Cohort 2</td>
<td>M</td>
<td>3.25</td>
</tr>
<tr>
<td>(N = 56)</td>
<td>SD</td>
<td>1.31</td>
</tr>
<tr>
<td>Cohort 3</td>
<td>M</td>
<td>2.94</td>
</tr>
<tr>
<td>(N = 68)</td>
<td>SD</td>
<td>1.52</td>
</tr>
<tr>
<td>Total</td>
<td>M</td>
<td>3.24</td>
</tr>
<tr>
<td>(N = 238)</td>
<td>SD</td>
<td>1.39</td>
</tr>
</tbody>
</table>

** $p < 0.01$. *** $p < 0.001$.  

324  Lai, McNaughton, Amituanai-Toloa, Turner, and Hsiao
Growth Modeling

In modeling the growth over time, we first theorized that on the basis of initial exploratory analyses of our data (Lai & McNaughton, 2008), we would improve achievement across the school year but that achievement would plateau (further progress would not occur) between school years. Hence a reasonable null model would be a four-stage model specified by the parameter vector $(m_1, m_2, m_3, m_4)$ indicated by Figure 6. (This is similar to other models tracking achievement over time, such as Borman, 2000.) When the full six-parameter model was tested against this null model, the null hypothesis was not rejected, $F(2, 1149) = 2.68, p = 0.07$.

We refined the model to give the most generalized representation, one in which the increase within a year was constant for all year levels. The fixed-effect part of this model was $\mu + \theta \times i$, where $i$ is the number of successive interventions that a student had experienced prior to taking the observation. The point estimate of intervention effect on achievement changes over time indicated that for every additional intervention a student received, the mean stanine level was raised by 0.42 ($SD = 0.04$), yielding a 95% confidence interval (CI) of 0.34, 0.49. This reduced generalized model estimates that each intervention raised the mean stanine level by between 0.34 and 0.49. The reduced generalized model was not statistically different from the first model, $F(4, 1149) = 2.07, p = 0.08$.

Hierarchical Linear Modeling (HLM) and Generalized Effects of the Intervention. The intervention was designed from the profiles of the local students and their instruction and contained elements that were designed to be both generic and personalized using cultural and linguistic resources. A telling indicator of the effectiveness of the intervention would be its generalized effectiveness. To further incorporate the notion of the reasonable null model with other factors and provide better modeling for the correlation structure of the repeated measures, HLM
was used to develop the growth curves. The Level 1 within-students’ correlations between repeated measures was fitted with an unstructured variance–covariance structure. This means that we made no assumptions regarding equal variances or correlations about the distribution of the measured data. The reasonable null model was refitted with the HLM structure, where this HLM version of the reasonable null model can be written mathematically as follows:

\[ y_{it} = \beta_0 + \beta_1 x_{1t} + e_{it} \quad (1), \]

where \( y_{it} \) is the stanine results for Student \( i \) at Time \( t \); \( \beta_0 \) is the Level 1 intercept, defined here as the expected achievement level; \( \beta_1 \) is the expected rate of increase in reading skills measured by stanine per each intervention received by Student \( i \); \( x_{1i} \) indicates the number of successive interventions Student \( i \) had experienced prior to Time \( t \); and is the within-student error of prediction for Student \( i \) at Time \( t \) with the variance of \( e_{it} = \Sigma \). The point estimate of intervention effect on reading achievement changes over time indicated that for every successive intervention a student received, the mean stanine level was raised by 0.30 \( (SE = 0.03; 95\% CI = 0.24, 0.35), t(237) = 9.86, p < 0.001 \). This model yielded an AIC of 4,273.4 and a BIC of 4,363.3. Thus, the model depiction in Figure 6 also applied to this HLM basic model. This level of gain is within the CIs of the original generalized model in Figure 6, establishing that it can be applied to the data. The slight differences in level of gain are due to the specificity of the HLM model to subgroups of students.
To improve the model fits further and provide generalized inferences to more specific student characteristics, we added two factors to the model specified by Equation 1 after a series of model selections. The two added factors combined examined the effectiveness of the intervention for different year levels and students with different initial reading achievement as well as the changes in their reading achievements over the summer holidays in greater detail. The final complex growth model can therefore be written as follows:

\[ y_{ijt} = \beta_{0ij} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + e_{it} \]  

where \( \beta_{0ij} \) is now the Level 1 intercept, definable by the expected achievement level for Cohort \( j \) to which Student \( i \) belongs, with the student's initial achievement level; \( \beta_1 \) is the expected rate of increase in reading skills (in stanine) per each intervention received by Student \( i \) of Cohort \( j \); \( x_{1ij} \) indicates the number of successive interventions Student \( i \) had experienced prior to Time \( t \); \( x_{2ij} \) indicates the number of successive summers that Student \( i \) of Cohort \( j \) has had since the first intervention, with \( \beta_2 \) being the expected rate of change in reading skills (in stanine) per each summer break for the respective cohort. The complex model yielded an AIC of 4,006.4 and a BIC of 4,079.3 and is a better model fit.

The estimates for the complex model with respective CIs are summarized in Table 7. Initially low-achieving students made more gain than did initially high-achieving students. For example, the point estimate for initially low-achieving students in Cohort 1 (Year 4 at Time 1) indicated that for every additional intervention experienced, their average stanine improved by 0.45; high-achieving students at the same cohort gained 0.22 of a stanine in their reading achievement. Similar patterns were found for the other two cohorts.

There was a differential effect on the students’ reading achievements over the summer holidays. For Cohort 1 (the largest cohort), regardless of initial reading achievement, the average stanine dropped by only 0.05 over the summer, consistent with the earlier plateau hypothesis. However, for Cohorts 2 and 3, the estimated stanine drops over the summer holidays were 0.54 and 0.24, respectively. There were no differences between students with different initial achievements over summer.

Throughout the model-selection process, gender and ethnicity (in terms of whether a student was New Zealand Māori or non-Māori) were not found to be significant effects in determining reading achievement levels, as including these factors did not improve the model fit. A series of statistical tests of significance indicated these factors were not statistically significant.

Discussion

The question asked in this article was whether a long-standing challenge for more effective teaching in a particular context of schools serving culturally and linguistically diverse poorer communities could be addressed. The challenge has been to accelerate levels of achievement in reading comprehension for Māori and Pasifika
students in Decile 1 urban schools in New Zealand. Low achievement in literacy has a long history, and there is some evidence that gains have occurred recently in beginning literacy instruction. But national and international evidence about New Zealand children suggest that large disparities in reading comprehension achievement remain (Crooks & Flockton, 2005; Flockton & Crooks, 2001). This pressing local problem has a more global significance, which is addressing gaps in reading comprehension for cultural and language minority students in schools serving the poorest communities (August & Shanahan, 2006; Sweet & Snow, 2003).

A proposed solution with two general components was tested within our research practice collaboration. We predicted that together the two components would effect changes in instruction that would result in changed achievement. We argue that although the two components draw on different theoretical traditions,
each are necessary to an effective solution, echoing in part an argument about thinking about both instruction and schools recently mounted by Pressley (2006) when identifying what the future of reading research could be. The components were the need to design effective instruction from contextualized evidence of teaching and learning and to systematically and collaboratively collect, analyze, and discuss evidence, which then is acted on to change practices. Other researchers have noted the assumed significance of the first component in comprehensive school reform (Borman, 2005), and research in the professional development literature also signals the significance of the second component (Timperley et al., 2007). These two general components were present in the three-phase collaboration implemented with the cluster of seven schools. The quasi-experimental design with its in-built replication and comparison with a matched cluster and controls for differential participant loss provides qualified support for the claim that the significant changes that did take place in reading comprehension achievement resulted from the intervention.

Here our concern was to test whether the intervention addressed the acceleration problem. Essentially, this is a benchmark for effectiveness. Did the intervention produce accelerated gains relative to expected gains, such that the distribution of students over a medium term of three years came to match a normal distribution? Moreover, was the intervention designed well enough so that there were generalized positive effects on particular groups of students who, under some circumstances, were likely to be differentially affected?

Gains did occur, and the evidence based on the quasi-experimental design format strongly suggests that the three-phase process was instrumental. The effect sizes associated with these gains were higher than those reported internationally for short-term gains in schooling improvement (effect sizes < 0.2) and were comparable and in some instances higher than those reported for sustained reforms over several years of about 0.5 (Borman, 2005). More important here was finding that significant acceleration in achievement had occurred. The students made approximately one year's gain in addition to expected progress over the three years. The generalized growth model fitted to the data suggested a rate of school year gain of between 0.30 and 0.50 stanine occurring year by year, with a plateau between school years. The students who had been on average in the below-average bands were after three years on average within the average bands, with the chi-square analysis indicating that the achievement distribution now closely resembled the national distribution, unlike at the beginning of the intervention. Seventy-one percent were now in middle to upper bands of reading comprehension for age level compared with only 40% at the start and 77% nationally, an educationally significant achievement given the long-standing nature of the challenge to effective instruction.

The improvements are even more significant considering the national picture of improvements in these age bands. There is no national testing in New Zealand in primary and middle school. However, a recent study examining trends in national databases for students of the same age range as those reported here revealed
that nationally, scores in reading comprehension have remained relatively stable for many years despite substantial changes in oral reading accuracy (Elley, 2005). A recent national review of all government-funded schooling improvement initiatives further indicates that only one initiative has been able to improve achievement for culturally and linguistically diverse communities (Annan, 2007). That initiative is the one we are reporting in this article. In other words, there is little evidence, either nationally in the age range on which we have focused or locally in interventions in the communities with which we are working, of a general improvement that could have contributed to the achievement gains seen here.

The intervention was designed from the profiles of the local students and their instruction and contained elements that were designed to be both generic and personalized using cultural and linguistic resources. A telling indicator of the effectiveness of the intervention would be its generalized effectiveness. The intervention showed generalized effects across gender and ethnicity. There were, however, some year-level variations, particularly over the summer holidays. Previous studies (e.g., Lai & McNaughton, 2008) have reported consistent year-level differences, particularly a drop in scores between Year 6 and Year 7 over the summer holidays unlike in other year levels, and have investigated several possibilities for the drop, including a change in the test format. However, the differences in year level in our study do not follow the same consistent pattern, and further investigation is required to understand these differences.

The analysis of growth over time suggests that change was not smoothly linear or quadratic. Rather, it had a more staircase shape, like a double sigmoid, brought about by the tendency for achievement to plateau over summer and for accelerated gains to be made during the school year. This model is similar to the one produced by Borman (2000), who noted a similar staircase shape for disadvantaged students, albeit with sharper drops over summer for students not attending summer school and a plateau with minor gains for students who had attended summer school. The summer effect, in which gains in school literacy, especially those of children in lower socioeconomic and minority communities, decrease from the end of one school year to the beginning of the next, has been related to family, social, and cultural practices that provide differential exposure to school-related literacy activities (Cooper et al., 2000). The generalized model developed from the data reported here showed a plateau rather than a drop over the summer, which may suggest that the family practices in New Zealand communities over the summer may have different features than those of the communities represented in data used by Borman and others. Anecdotal reports from the present schools implicate the importance of local community libraries and of encouragement of students’ love of recreational reading as influential over summer.

The nonlinearity of growth across successive years is consistent with a model of reading in which development in the middle school years was constructed through practices both in school and out of school. To better understand how relationships between school and home over summer determine patterns of progress, we clearly need to add an ecological dimension to the model of reading.
development (Bronfenbrenner, 1979). Previous studies have identified the importance of continued access to and engagement in school-like literacy practices in family and neighborhood contexts, such as recreational reading (Anderson, Wilson, & Fielding, 1988) and practices described in the classic longitudinal study by Heyns (1978). The notion of increasing independent engagement by the middle year needs to be tempered with how that increasing independence is constructed or supported by practices and resources both at school and at home. The development processes suggest the likelihood that both schools and families (Anderson et al., 1988) may be sites at which to directly measure the probability of access to and use of specific practices. The presence of variations in growth patterns across cohorts and at different times clearly indicates the potential for development changes.

The intervention produced gains during the school year that were substantial enough to overcome the plateau associated with the summer holidays. Increasingly higher levels were achieved by the end of each successive year, meeting the tough challenge for judging effectiveness. We have not analyzed the phase-by-phase data separately here, but we are currently analyzing the relationships between gains during the year and over the summer and relating these to reports from the schools about community literacy activities over the summer. What is apparent is that the drop is not inevitable.

These data also signal the importance of including the summer break when modeling growth across school years to gauge implementation effectiveness. It is possible to test at a single point in the year over multiple years, thereby reducing the unequal time intervals, accounting for the effects of the summer, and enabling a different model to be fitted to the data. However, this ignores an important source of information about the changes between school years that affect literacy growth over time. It may also conflate the efficacy of the intervention with possible community, social, and family effects over summer. The need to understand this variation further and specifically to consider the role of the school in promoting community and family literacy activities is signaled in these data (Cooper et al., 2000). This is all the more pressing given that the data suggest that, by and large, the summer drop affects students of different genders and ethnicities and students from low- and high-starting achievement levels in similar ways.

Both the general outcomes and the growth curve also show the need for continued gains because the match with the normal distribution was not perfect, and the predicted time to reach normal levels was longer than three years. Extrapolating from the fitted model, we would need another two to three years of sustained acceleration, given a continuing plateau over each summer. Borman’s (2005) summary of general effects for schooling improvement suggests modest gains in a first year may be followed by an implementation dip and that greater gains accrue over the medium to longer term, around five to seven years. One implication of these findings is a view of research and development collaborations in which sustainability is built into the collaboration rather than assuming that it will occur independently.
The intervention was based on two components that have been described and analyzed fully in other publications. But it is worth noting that our results support the predicted significance of both components. Other studies have linked the analysis of data to improving achievement (e.g., Cawelti & Protheroe, 2001), and Hawley and Valli’s (1999) review of professional development identifies problem solving around evidence gathered from one’s own school as a more effective form of professional development than traditional workshop models. Further details on how such a process can be developed and what it should look like are contained in Robinson and Lai (2006), who presented the methodology underpinning the analysis processes used in this study and provided 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 as well as drawing valid inferences 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.

What this suggests is that, in general, the teachers through the professional communities within schools had the capacity to change practices but needed support to identify the locus of change and test their theories about raising achievement. Given the close collaboration with researchers, this also confirms the importance of external support in particular research–practice collaborations (e.g., Annan, 2007; Robinson & Lai, 2006). However, identifying collaborative problem solving as important is not an argument against the need for specificity in interventions. In many respects, the analysis process was quite detailed and specific around the hypotheses for more effective instruction. Moreover, the level of detail required to fine-tune strategy instruction, for example, meant that the teachers were required to act more like the “adaptive experts” described by Bransford et al. (2005, p. 48), with detailed pedagogical content knowledge.

Moreover, the complexities in our data around change over time in subgroups 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. A number of research programs (e.g., Taylor et al., 2005) use some combination of teacher analysis of data 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.

The second component of the intervention involved the targeting of specific strengths and needs, using evidence from the local achievement data and teaching practices. The need to contextualize the evidence is both to better inform problem solving and to add to theoretical understandings of instruction in cultural and linguistically diverse contexts (Garcia, 2003; Sweet & Snow, 2003). The analysis of profiles of teaching and learning provided specific instructional hypotheses. It showed that low decoding levels were generally not a problem. Instead, patterns
of checking and detecting threats to meaning in paragraph comprehension, the size and knowledge of vocabulary, incorporation of students’ familiar knowledge complemented by building students’ awareness, and finally the density of instruction were areas of concern. The four areas to enhance were identified through the problem analysis process and then were targeted more specifically in the professional development workshops in the second phase. The evidence from the measures of integrity was that the focus and fine-tuning in practices occurred around these four areas of concern, although there were still low levels of use of cultural and linguistic resources (McNaughton, MacDonald, Amituanai-Toloa, Lai, & Farry, 2006).

An example of how research grounded in the contextualized problem solving can contribute to our understanding of teaching and learning can be found in the analyses of strategy instruction. A solid research base provides considerable evidence for the significance of developing comprehension strategies (Pressley, 2002). But what was not initially anticipated from that research base was the specific problem with strategy instruction that we found in this context. Having reexamined the literature, we note that previous commentators had signaled the same problem with strategy instruction. The problem was repeated explicit instruction out of the context of reading connected text for meaning (Baker, 2002; Moats, 2004).

The problem is likely derived from the tendency for instructional packages to be presented and then 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 of enabling readers to construct and use appropriate meanings from texts (Pressley, 2002). The increased focus on checking and building awareness of use over the intervention was associated with the gains in component tests, including paragraph comprehension (McNaughton et al., 2008). 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 in our context.

The solution to this risk lay in the collective evidence-based problem solving and the increased knowledge that teachers developed to understand the nature of comprehending, learning, and teaching and the characteristics of effective teaching. These are features of effective programs that have been identified by other researchers too (Taylor et al., 2005). More generally, this carries implications for the features of effective teacher education and professional development. The issue here is the balance between teachers learning and carrying out predetermined patterns of instruction known to be effective or developing as adaptive experts with a body of knowledge and practices who can use and modify known instructional practices to solve issues of effective practice (Bransford et al., 2005; Robinson & Lai, 2006).

The three-phase model for the intervention was implemented readily in the schools, as indicated by the take-up of the problem solving in and across schools in the first year. Successful implementation of the model was also indicated by
no noticeable increase in teacher turnover and no school dropout as in the early stages of some of the school reform programs (Slavin & Madden, 2001). We attribute this to two features of the context. One is the general New Zealand context, where historically there has been problem solving and innovation at the local level. One example is the development of the “natural language” text, primarily by teachers dissatisfied with texts with little relevance to or familiarity with local children (McNaughton, 2002). A second feature is that our research team had been working with these schools on related projects for five years prior to this study (Phillips et al., 2004), and considerable mutual trust and protocols for collaboration had developed, a feature of effective longer term school reform efforts (Taylor et al., 2005).

The data we have examined here were longitudinal, which meant that the analyses were restricted to just those students who were present over three years. A significant issue facing the sustainability of schooling improvement projects is to test the impact on new cohorts of students and those students who are present intermittently over a period of time against the acceleration criterion. Analyses will need to be conducted to check the applicability of the model of growth for new and transient students to make further judgments about the effectiveness.

Another significant issue is scaling up the intervention to different schools and different contexts. Our intervention model is an alternative to prescribed instructional programs, based on a specified process of contextualizing instruction to unique profiles of needs and analyzing and discussing these profiles in professional learning communities. Although successful in accelerating achievement, the difficulty with scaling up comes from guaranteeing fidelity. If fidelity is to the content of the instructional program, then scaling up is not possible, as the program of instruction may differ across contexts due to the unique profile of needs. However, if fidelity is to the process, then scaling up requires replicating the specified process, which may result in different instructional programs appropriate to local needs. We have been engaged in replicating the process, scaling up this intervention to like schools (similar socioeconomic and ethnic groups with low-starting achievement) and unlike schools (different socioeconomic groups and ethnic groups with average-starting achievement). In the replication series, we have found that achievement has accelerated across the 40 schools examined (Lai, McNaughton, & Hsiao, 2008).

In conclusion, we revisit the primary question asked in this article: whether a long-standing challenge for more effective teaching in a particular context of cultural and linguistic diversity could be addressed. The results from our study provide further evidence that the low achievement results for such students are neither inevitable nor immutable. Rather, with effective instruction from contextualized evidence of teaching and learning and with systemic collection, analysis, and discussion of evidence that is acted on to change practices, culturally and linguistically diverse students can succeed. Most important, the acceleration criterion enables us to better judge what it will take and how long it will take to achieve equitable outcomes for all our students.
1. Why is the disparity of reading comprehension achievement a problem of global significance?

2. How can educators design interventions that address the “summer slumps” in reading achievement?

3. Why is the collaborative model that the researchers implemented in this study an example of theory generation and revision?

4. How can researchers assist and support teachers who work with language-minority high-poverty urban students?

NOTES

*When this chapter was written, Turner was in the Starpath Project at the University of Auckland.

The Woolf Fisher Research Centre receives support from the Woolf Fisher Trust, the University of Auckland, and Manukau Institute of Technology. The research reported here was funded by the Teaching and Learning Research Initiative (New Zealand Council for Educational Research), the Ministry of Education and the Woolf Fisher Trust. We thank all the schools, teachers, and children who took part in this research.

REFERENCES


New Zealand: University of Waikato, Education Department.


