MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

THE ASSESSMENT OF TEACHER QUALITY

AN INVESTIGATION INTO CURRENT ISSUES IN EVALUATING AND REWARDING TEACHERS

EDUCATION POLICY RESPONSE GROUP
INSTITUTE OF EDUCATION
MASSEY UNIVERSITY

SEPTEMBER 2013

Ivan Snook
John O'Neill
Stuart Birks
John Church
Peter Rawlins

S. M. Richards, 2003

THE ASSESSMENT OF TEACHER QUALITY

An investigation into current issues in evaluating and rewarding teachers

EDUCATION POLICY RESPONSE GROUP

Institute of Education

Massey University

Ivan Snook

John O'Neill

Stuart Birks

John Church

Peter Rawlins

September 2013

CONTENTS

SUMMARY

1. In the past thirty years a movement for educational reform has spread throughout the world. Now known popularly as the Global Educational Reform Movement (GERM) it is promoted by international development agencies and private enterprises seeking to influence educational policy. This movement advocates competition, choice, business-style management, curriculum standardization, "value added" assessment of teacher effectiveness, and performance pay.

2. When we look at recent developments in New Zealand education and policy documents emanating from the Treasury it seems clear that many elements of GERM are on the ascendancy here. Viewed in this way, many particular innovations (such as national standards, charter schools, the extension of integrated status to elite private schools, and the proposals for schools in Christchurch) seem to fit neatly into it.

3. The Treasury policy agenda would depart significantly from the social democratic philosophy underlying the reforms of *Tomorrow's Schools*. It seeks to initiate a regime in which educational research (however controversial, subjective and unreliable) is to be interpreted by state officials and used by state agencies to determine in close detail what happens daily in all state classrooms and, hence, influence very significantly the educational experience of every child. It is very important, therefore, that these proposals and the claims which underpin them should be subjected to critical scrutiny.

The Treasury policy agenda (Section 1)

4. Treasury's education policy agenda and advice to government are based on conventional economic theories of business process and human capital. In these it is assumed that variations in the quality of teachers and teaching can be removed over time through the use of performance incentives based on value added student achievement data. This, in turn, will resolve the problem of structural education inequalities.

5. Treasury's analysis is simplistic and largely ignores or discounts other plausible contributory causes to educational inequalities. For example: the education system's existing policy settings may not support teachers to prepare students for assessments; some parents may spend considerably more than others in supporting their children's education outside school; increases in government funding may not keep pace with demands made on the school system itself; or inequalities in achievement may be a function of social and economic inequalities outside the school and outside the teacher's control.

6. It is also assumed that value added assessment measures and performance-related pay incentives will: (1) be valid, reliable and practicable; (2) be motivational to teachers; and (3) produce no unintended consequences. Such assumptions may be wrong and need to be subjected to critical educational scrutiny before such a radical departure from current policy settings is countenanced for New Zealand children and teachers.

Effective teaching: an international perspective (Section 2)

7. A 2005 OECD study of 25 educational systems around the world produced three main findings: (1) the largest variations in student learning are attributable to what students bring to school (ability, attitude, background, and so on); (2) "teacher quality" is the single most important school variable; (3) Most research on teacher quality is narrowly focused on test scores and readily measurable teacher characteristics. Characteristics that are harder to measure, but may be just as vital to student learning (such as clarity, enthusiasm, creativity, warmth and the ability to create effective learning environments and relations), are typically not measured.

8. Despite the overwhelming findings about the effects of home background, politicians and the popular media persist in minimising it and insisting that great improvement in learning can be gained by concentrating on schools and teachers in isolation from home background and parental income.

9. Numerous studies demonstrate that the socio-economic background and prior experience of students are the main influences on learning. Their effects need to be carefully taken into account if they are not to be overlooked in the creation of school-

based “solutions” which cannot work and which merely alienate teachers whose morale and cooperation are essential.

10. Despite the OECD’s claim of the critical importance of teachers, it is important to remember that, in another way, the student is of most importance: what he or she brings to school in the way of personal ability, cultural norms, values, skills, and background experience plays a central role in what and how he or she learns.

International studies of student achievement (Section 3)

11. Treasury (and Government speakers) frequently refer to international studies of school achievement to indicate that all is not well with the New Zealand education system. In particular they claim that: (1) New Zealand has a wider distribution of student achievement than similarly achieving countries, in other words “The long tail of underachievement”; and (2) that due to this wide distribution of achievement, New Zealand’s education system is considered only “fair to good”.

12. We examine in detail the findings of PISA (which are most often quoted by Treasury and Ministry speakers) and also look at TIMSS, PIRLS and our own New Zealand based monitoring study NEMP. Overall, the results do not support the Treasury’s claim that New Zealand has a “long tail of underachievement” in relation to similarly achieving countries.

13. On PISA New Zealand has both a higher than average proportion of students in the higher proficiency levels and a lower than average proportion in the lower levels on tests of reading, mathematics, and science literacy. This indicates that, although New Zealand might have a wide spread of achievement, this spread is situated towards the upper end of the achievement spectrum. High achieving students are doing better and lower achieving students are generally performing on a par with comparable countries.

14. On the other hand, the TIMSS data tell us that, compared to other similar countries/economies, our performance in mathematics and science is rather poor at Year 5. However, students improve as they get older, in both their average performance, and in the proportion of students meeting international benchmarks. The

PIRLS data tell us that there is a relatively wide spread of achievement in reading at Yr 5 with higher proportions of lower achieving students than the international median. According to PISA, however, by age 15, our lower end performance is similar to comparable countries and large proportions of our students reach higher benchmarks.

15. Treasury also claims that New Zealand teachers are less successful than others in counteracting socio-economic disadvantage but fails to acknowledge that in the late 2000s, New Zealand was one of the most unequal countries in terms of income distribution (Gini index), ranking 28th out of 31 OECD countries. Between the mid 1980s and mid 2000s, income inequality rose in 15 of the 19 OECD countries for which data are available. New Zealand (together with Sweden) had the highest cumulative rate of change in this period. In the same period, there were significant increases in the numbers of people living in poverty (below 50% of median income) in New Zealand.

16. Predictably, PISA found that the impact of economic, social and cultural status on students' performance in New Zealand is the highest among the OECD countries and the likelihood of disadvantaged students performing at levels similar to those of their advantaged peers is much lower in New Zealand than in countries such as Australia, Canada, and Japan.

17. However, despite this, the proportion of "resilient" students—those who come from a disadvantaged socio-economic background and perform much higher than would be predicted by their background—is 9.2% of all New Zealand students, statistically higher than the OECD average of 7.7%. Moreover, if we take this as a proportion of those who are considered disadvantaged then approximately 36% of disadvantaged students can be considered "resilient", higher than the OECD average of 31%.

18. The interpretation placed on these data by the Treasury is that New Zealand teachers have been less successful than others in overcoming social disadvantage. This is surely to misunderstand the PISA data which, on the contrary, suggest that despite higher than average levels of income inequality and child poverty in New Zealand society, higher than average numbers of students overcome the odds and achieve.

19. Although some international comparison studies might identify areas of content or skills where students are relatively stronger or weaker nationally, this could equally well be related to the structure and content of the curriculum or preferred pedagogy. In many instances these are national policy mandates rather than teacher or school decisions. Accordingly, in order to provide a balanced picture, these studies should be considered in context, in their entirety, along with other national measures of student achievement, in light of the international research literature, and in consultation with expert educational groups.

20. The Treasury claims that the third McKinsey Report shows that as a result of our wide distribution of scores, the New Zealand system is rated only "fair to good". Inspection of the report reveals that although it investigated 20 educational systems from around the world in some depth, New Zealand was not among them. Indeed, New Zealand is mentioned only twice in the report, once on a graph depicting various countries' achievement levels in international assessments against public expenditure per student, and once in the list of OECD countries that act as a control group in developing a "Universal Scale". Even if we accept New Zealand's position on this Universal Scale, the claim that it is our wide distribution of achievement scores that has placed us in a particular position on the scale is not supported by the methodology used to create it.

21. On our analysis, the Treasury's principal claims about New Zealand's performance in international assessments are false: (1) There is no "long tail of underachievement" in any general sense; (2) New Zealand teachers are not deficient in remedying socio-economic disadvantage; and (3) The claim that New Zealand education is only "fair to good" because of its "tail" of underachievement, is unsubstantiated by the report which is used to support the claim.

Teacher effectiveness (Section 4)

22. While officials and politicians favour the term "teacher quality", there is no agreed definition of what this means in practice, other than raising student achievement scores. Research studies of teacher effectiveness focus on what the teacher does and what the student learns. All of the studies we reviewed have limitations and therefore suggest caution on the part of policymakers who wish to link periodic assessments of

teacher effectiveness with teacher registration, employment, career progression and/or remuneration.

23. These studies generally have been limited to measures of improvement in reading (and sometimes mathematics) and have largely ignored all other schooling outcomes such as improvements in the other six curriculum areas, improvements in social skills, improvements in self-directed ability to learn and improvements in the unique talents of individual children.

24. All of these studies leave a major part of the variation in student achievement unexplained. This is because most of the measures of achievement used are not very accurate and because many of the factors affecting achievement (e.g. home background) have been measured with even less accuracy.

Value added measurement (Section 5)

25. In recognition that teachers cannot be held accountable for effects outside their control (e.g. home background, previous knowledge, classroom composition, natural ability and mental engagement) researchers have turned to trying to assess what "value" a teacher has "added" to the student's achievement over the time they were in his or her classroom. This is a particular form of achievement based assessment known as value added measurement (VAM). VAM is problematic because: (1) it takes little account of differential student abilities; (2) it is difficult to control for the widely different learning experienced by students outside school; (3) the variables which are the most easily rated are not necessarily the best for assessing the overall quality of learning; and (4) there are many other explanations for improved achievement scores.

26. VAM scores have been shown to be very unstable, with dramatic fluctuations from year to year, and high rates of measurement error. Consequently, the scores gained by teachers in the first year of assessment, cannot predict even broadly their scores in the second and subsequent years. VAM schemes fail partly because so many factors influence student learning. Among these are: (1) other teachers of the child; (2) the quality of school resources; (3) the variety of out of school learning experiences at

home and in the community; and (4) the 'learning loss' that occurs over the long vacation. These factors do not influence all students equally.

27. There are also negative consequences that flow from reliance on VAM. The main ones are: (1) it leads to narrowing of the curriculum; (2) teachers are discouraged from working with special needs students and children from low income homes, and schools (anxious for their reputation) resort to all kinds of strategies to exclude such children; and (3) teachers are discouraged from co-operating with other teachers on the grounds that they might thereby lose some tactical advantage. At a time when team teaching and shared professional knowledge are regarded as educationally very important, VAM works against the cooperative spirit.

28. Econometric studies suggest important connections between teacher qualities and student achievement on tests but they face serious statistical and analytical problems and are unable to provide any analysis of what constitutes a good teacher or any advice to policy makers regarding teacher recruitment, training or reward. Based on documented error rates in such approaches, we estimate that if there were 100,000 teachers in a system and three years of student achievement data were used, at least 5,000 teachers would be dismissed unfairly and 5,000 underperforming teachers would be retained. If the administrators used data accumulated over 10 years, "only" about 2,500 teachers would be wrongly categorised. The suggestion, therefore, that teachers be dismissed on the basis of student achievement is simply bizarre

High stakes teacher assessment (Section 6)

29. Two specific teacher performance evaluation schemes were examined, the Teacher Evaluation System (TES) in Cincinnati, USA, and the Contextual Value Added (CVA) scheme in the UK. The notion of "performance" in each is quite distinct. In the TES, the teacher is evaluated on what she does (her performance, like an actor). In the CVA, the teacher is evaluated in terms of "how successful she is" (her performance in raising the "achievement" of her students). This dichotomy is marked by different approaches to the measurement of "performance".

30. The Cincinnati TES does not involve any measurement of student learning and hence is not a "value added" tool. It is a "standards" based procedure which is intended to examine and refine the activities of teachers through multiple structured classroom observations by trained evaluators who are experienced classroom teachers. Comprehensive evaluations are undertaken at approximately five yearly intervals for experienced teachers and more frequently for newer teachers. In addition, a scheme is in place to assist new and experienced teachers who experience difficulties. Two studies we reviewed have claimed to identify increases in student achievement as a result of the scheme's introduction. However, the reported student achievement gains have been modest, inconsistent across learning areas, and could equally well be explained by other variables not considered by the study authors.

31. The English CVA sets out to show the progress children have made while attending a particular school. Unlike statistics that merely report actual exam performance, the Contextual Value Added system attempts to take into account the circumstances of children attending the school that are beyond the school's control. This is done by comparing a child's performance with that of children with a similar prior performance and similar circumstances. Numerous variables are included in the model in an attempt to "control" for factors beyond the school or teacher's control. Conservative politicians have claimed that the statistical methodology discriminates against socially advantaged schools, while liberal/progressive politicians have claimed that it underestimates the challenges faced by disadvantaged schools. Researchers have claimed that the CVA suffers from serious reliability and measurement errors to the extent that the results are only marginally more reliable than unadjusted achievement test results.

Performance pay schemes (Section 7)

32. One major review of performance pay concluded that: (1) judgments about teacher performance for "high stakes" purposes such as registration, reward, career advancement and promotion must be based on several sources of evidence; (2) performance standards are multidimensional (e.g. knowledge, skills, dispositions), and therefore multiple forms as well as multiple sources of evidence are required; and (3) valid and reliable assessment results require independent, trained assessors together

with evidence about the context in which judgments about performance are being made.

33. Performance pay schemes that are linked with student achievement adopt two quite distinct approaches: (1) teacher performance is based on student performance on standardised tests; or (2) teacher performance is based on evidence about what students are doing in classrooms as a result of conditions for learning established by teachers.

34. In our view teachers may be held responsible only for exercising their professional judgment to create supportive conditions for students to learn: they cannot be held responsible for what, or how much, students learn since this is beyond their control.

35. Rewards can only motivate additional effort. They cannot teach new skills. Rewards will motivate additional effort only if the recipient judges that they are worth the extra effort. Rewards can have perverse effects. For example they can function to reduce intrinsic motivation. Given that much teacher effort is intrinsically motivated, great care must be taken to avoid any kind of extrinsic reward system that would damage this intrinsic motivation.

Conclusion (Section 8)

36. We conclude by noting that Treasury's policy advice for education is based on a relatively simple economic model of the effects of investing in human capital: variations in teacher quality strongly influence variations in student achievement. To secure the required improvements in teacher quality, incentives and sanctions should be applied according to gains or losses in expected student achievement.

37. We have already argued that teachers are only one of many factors that influence student achievement and that Value Added Measures are entirely unreliable. Therefore they cannot ethically be used to reward or sanction teachers. Nevertheless, we strongly support the goal of improving school effectiveness.

38. We agree with the McKinsey and Company's common sense observation that at system level: "High performing school systems consistently do three things well. They

get the right people to become teachers (the quality of an education system cannot exceed the quality of its teachers). They develop these people into effective instructors (the only way to improve education is to improve instruction). They put in place systems and targeted support to ensure that every child is able to benefit from excellent instruction (the only way for the system to reach the highest performance is to raise the standard of every student".

39. We are also aware of the OECD's observation that teaching profession is in long-term decline, yet there is increased awareness of how crucial teachers can be in the achievement of students and in the progress of society. It is essential, therefore, that we avoid policies (such as constant surveillance and performance management) which alienate teachers, deter students (especially the most gifted) from a teaching career, encourage unprofessional behaviour (such as cheating on tests) and lead to rapid turnover of teachers in a system that needs stability.

40. The Treasury agenda for education is a dangerously narrow and simplistic interpretation of what the evidence says about assessing and improving the quality of teaching. Bluntly, their solution is the problem in our view.

41. Instead of Treasury's business process model and economic theories, we urge consideration of a range of policy options that are based on educational theories and research evidence that recognises the complexity of teaching and learning and the social contexts in which they occur both in and out of school.

42. Broadly, at the system level we argue for policy settings that place higher trust in teachers and in the power of teacher collegiality to improve teaching; for targeted student engagement and retention policies in the middle and senior secondary years, particularly in low decile schools; and abandonment of the current, harmful obsession with National Standards and NCEA Level 2 targets.

43. We observe that various policy and funding decisions over the past twenty years have led to a situation in which many candidates for initial teacher education in the primary school sector may not have the knowledge and skills needed to deliver the mathematics and science curricula. If this is true, simply measuring student achievement and

sanctioning teachers will do nothing to address underachievement. Indeed, in this context, proposals to reduce teacher education programmes from three or four years to one year by making primary ITE postgraduate entry only are of considerable concern.

44. We offer six alternative policies that, if implemented carefully with sufficient resources, have the potential to make schools more effective than they are currently.

45. Teachers cannot change the things they do not know how to do. For a teacher to become more effective, that teacher must acquire new skills. This requires the teacher to discover that a more effective teaching method exists, engage in active rehearsal of the new skill with feedback and to continue practising until they have achieved mastery.

46. Priority should be given to the acquisition by teachers of the new diagnostic, teaching and evaluation skills that have been shown through research to be more effective in fostering student learning than those which are currently being used.

47. Policies are also needed to accelerate the production of teaching materials that have been field-tested and shown to be effective in developing the learning outcomes contained in the New Zealand curriculum.

48. Policies are needed to upgrade the management, appraisal and professional development skills of all staff involved in education at system level and at school level.

49. All those involved in the system also need to develop clearer understandings of what counts as sound educational research and reliable research evidence, and how research may be used to improve teaching and learning.

50. Finally, education policies must encourage collaboration by capitalizing on the intrinsic motivation of school leaders and teachers. Motivation cannot be mandated and trust cannot be legislated for. Yet both are integral to securing system-wide improvements in schooling over the longer term.

We hope that we have pointed the way toward a better approach to that currently being advocated. We argue that alternative policies are required to enhance and upgrade the status of the teaching profession, to attract and retain high achieving candidates and to allow them to flourish as autonomous professionals. Although we do not claim to have all the answers, we submit that the evidence we have presented strongly suggests that the agenda we put forward is more defensible than that of the New Zealand Treasury and we welcome informed debate on it.

SECTION 1. THE NEW ZEALAND POLICY AGENDA

> *We have witnessed, over the past half-century or so, determined efforts to find general solutions to social problems...The cost to national economies has been prodigious, and there is precious little to show for it...It ought to be apparent by now that generalized, one-size-fits-all solutions do not work...* (Stringer, 2007, p. ix).

In the past thirty years a movement for educational reform has swept through many education systems, including the USA, England and Australia. Now popularly known as the Global Educational Reform Movement or GERM (Sahlberg, 2011), it is vigorously promoted by international quasi- and non- government organizations and private enterprises. These groups exert growing influence in public education policy development and enactment. In this movement the major emphasis is on: competition and choice, the standardization of curricula, the use of corporate management models, standardised measurement of teachers (by "value added" measures), and performance pay.

When we reflect on how educational policy has been developing in recent years and study various discussion and briefing documents, it seems clear that there are plans for major reforms to the New Zealand schooling system. When we view these plans in their entirety – "the policy agenda" – many specific initiatives seem to fit together to form a clear pattern. The initiatives include National Standards, a progress and consistency tool (PaCT), Partnership Schools Kura Hourua (charter schools), the extension of integrated school status to elite private schools, increasing the government funding subsidy to private schools, and the proposals to "restore", "consolidate" or "rejuvenate" schools in Christchurch.

1. The Treasury Agenda

In broad outline the agenda goes like this:

1. Treasury (2012) argues that economic growth (expressed in terms of per capita GDP) is dependent on increased educational achievement. This is sometimes based on an observed relationship between per capita income and some education measure, using aggregate country data for cross-country comparisons, although such relationships are

often very loose. It also can be anticipated from a production function approach to GDP, where aggregate output is a function of the quantity and quality of inputs. Labour is one of the inputs, and labour quality can be increased with suitable education. The point is then made that we can (and must) improve educational achievement and then we will get increases in GDP.

2. The claim is made that a significant a proportion of New Zealand students fail educationally. The main evidence for this, according to the Treasury is that:
   (i) Performance of New Zealand students in international tests has remained static over the past decade, despite a large increase in educational funding.
   (ii) New Zealand has a wide distribution of educational achievement and more low achieving students compared to other countries with a similarly high average score in international tests. (As it is commonly put, we have "a long tail of underachievement".)
   (iii) Among OECD countries, New Zealand has the largest variation in student achievement within schools (Treasury, 2012, pp. 1-2).
   (iv) The Treasury acknowledges that "Research on student learning consistently shows that the largest source of variation in school learning is attributable to differences in what students bring to school—their abilities and attitudes, and family and community background" (2012, p.2) but it goes on to assert that the main "in school" influence on student learning is the teacher: "Of those variables which are potentially open to influence in educational settings, factors to do with teachers and teaching are the most important influences on student learning" (p. 2). (See Appendix A.)

3. It follows that the key to improvement in education is to improve the quality of the teachers. "A recent report by Australia's Grattan Institute highlights how four East Asian countries have achieved significant improvements in the performance and equity of their schooling systems by building teacher capacity. They have done so via a focus on high quality initial teacher education, improved feedback and mentoring, and career structures that value good teaching". Therefore, "we need a system that selects the right people into the workforce, trains them appropriately, develops them as professionals over time, and prepares some of them for the challenges of leadership" (Treasury, 2011b, p. 3). The agenda to improve the quality of teaching in New Zealand requires

deliberate policy changes of which the major changes are: A better use of resources to rejuvenate the system; the introduction of more "competitive tensions"; the measurement of teachers in terms of the "value" that they add to student learning; and performance rewards to teachers and principals whose value added scores merit it.

The first claim about the connection between education and economic growth is a major topic in itself which we will not traverse here. It seems clear to us that the data regarding the relationship are capable of many interpretations. Though there is certainly a correlation, the causal direction is far from obvious. It is plausible to suggest that societies with higher incomes are able to afford more education (as they are able to afford more motorcars and telephones) and hence the argument that more education *causes* higher GDP is suspect. In addition to that, the correlation itself drops significantly when lower income countries are excluded. As Birks and colleagues put it, "The relationship between education and income is much weaker for higher income countries alone" (Birks, Snook, Prochnow, Rawlins & O'Neill, 2013, p. 5).[1] Thus it seems that once a country's economy reaches a certain level the pay-off from extending education is less clear.

We will simply assume that there is a good deal of doubt about the accuracy of this basic rationale for change and concentrate on "unpacking" the assumptions behind Treasury's position on the relationships between the quality of teaching and the quality of student outcomes. In particular we will demonstrate that the assessment of teacher quality is much more complex than the Treasury acknowledges, that value added approaches to teacher assessment are unreliable and that performance pay based on them would lead to huge injustices which would undermine teacher morale (which is widely recognised as centrally important for successful education systems).

In March 2012, Treasury released an "evidence brief" (Treasury 2012) to support the advice on education policy contained in its Briefing to the Incoming Minister of Finance (BIM) prepared during the 2011 election (Treasury, 2011b). Effectively, the BIM advice had asserted that the schooling outcomes of underachieving student groups in New Zealand, as measured by international test results, could be raised by improving the quality of teaching

[1] This paper can be viewed as complementary to the current paper, providing in some depth the economic arguments for some of our conclusions.

through the systematic use of value added data and further professionalizing the workforce (Treasury, 2012, p. 1). The six page evidence brief elaborated the major claims of BIM advice and cited 17 published sources to support them.

2. The Treasury Advice

The Treasury's advice can be summarised as follows:

- There have been large real funding increases for the schooling system in recent years.
- Yet, achievement levels for some groups (particularly Māori and Pasifika) remain unacceptably low.
- Within schools, low student achievement can be raised most by improving the quality of teaching.
- The quality of teaching can best be improved by more systematic use of value add data
- Increasing student/teacher ratios and closing some schools can free up funding to support these improvements in teacher quality.

1. There is some doubt as to the extent that real funding has increased as argued. Even if it has it may have: (a) increased during the years cited but not the years before or after this time span. (i.e. using a longer time span may not show a significant increase. Another possibility is that it has increased at a lesser rate than other test participating countries. If the average increase in other countries is greater than the New Zealand increase, New Zealand's expenditure could be said to have *decreased* in relative terms. A third possibility is that it has increased but not yet reached the funding levels of those other countries. This indeed seems to be the case: while New Zealand's expenditure has increased, it remains below average for OECD countries (Education Counts 2012). Judgments about system effectiveness and efficiency, overall and with sub-populations, must therefore be made with these caveats in mind (see Section 4).

2. It is wrong to imply that significant increases in funding should automatically be expected to lead to significant gains in student achievement outcomes, either for all students or for particular groups. First, as noted, the fact that funding may have increased does not account for the effects of family support and policy settings. Many things

contribute to students' educational outcomes, not just school funding levels. Second, general increases in funding may be needed just to maintain existing levels of inputs and outcomes. Third, incremental increases in funding may be enough to produce modest incremental increases in average student performance but be quite insufficient to address deeper seated challenges of entrenched under-achievement among particular sub-groups.

3. Similarly, it may be that the causes of educational underachievement may not be amenable to solution using general funding increases. In order to address a problem of educational underachievement, it is necessary to identify the problem, isolate its causes and address them directly and appropriately. So, for example, if general educational underachievement remains largely concentrated among the poor (and Māori and Pasifika are over-represented among the poor) this calls for quite a different policy response than if the problem of educational underachievement is adolescent girls' reading performance on a particular literacy test or their general understanding of and enjoyment of science.

4. It is problematic to relate differences in achievement solely to the quality of teaching. For example, before they begin to attend school, children spend five years developing cognitively and affectively in a wide variety of family and local community settings, and have had quite different amounts, types and quality of early childhood education. Children do not arrive at school as either blank slates or pre-programmed machines, or with identical knowledge, skills and attitudes to one another, nor are they fully malleable or suggestible by their teachers. Similarly, as they progress through the compulsory education system, children develop cognitively and affectively in diverse learning areas at different times and rates. Moreover, as they develop, their disposition toward school may vary according to the relative influence of family, classmates and friends, while non-formal and informal activities in their out-of-school lives may compete for their time, energy and commitment to formal learning in school. Logically, therefore, what children bring to the teaching-learning relationship is as important as what the teacher brings. Thus, one may reasonably say that only when all these potential differences have been properly weighed and taken into account could one plausibly claim that the quality of teaching makes the most difference to students' learning (and then only during their school years).

5. The concept of "value added measures" of teaching may have emotive appeal but it lacks precision. It is one thing to assert that the contribution of a teacher to a student's learning should be systematically measured, it is quite another to do so accurately and fairly. We examine Value Added Measures (VAM) in Section 5.

6. Playing off some policies (larger class sizes) against improving teachers fails to recognise the important interactions between such policies. While reducing class sizes may not of itself lead to improved teaching it can be argued that small classes make it *possible* to institute features (such as individual attention and feedback) which research suggests *are* important for improved teaching. Nor does "closing or amalgamating" schools take into account what is known about the importance of teacher-parent relationships or the necessity for there to be a good "mix" of students to enable the weaker to progress by association with those who are higher achieving. Bureaucratic reorganisation of schools seldom takes into account what is known from research about learning and teaching.

The agenda of Treasury (and presumably the Government) marks a new stage in New Zealand education. Historically the state has had an extensive but well contained role: the funding, provision and regulation of schools and the provision of general guidelines regarding the curriculum. The rest has been left to parental preferences, student needs and teacher professionalism. In 1989 the system was further decentralised to encourage more local initiative, with parents and teachers working in partnership to relate general guidelines to local conditions. Now, however, the government seems determined to overturn much of this decentralisation and to move beyond its traditional roles. It intends to initiate a regime in which educational research (however controversial, subjective and unreliable) is to be interpreted by state officials and used by state agencies to determine in close detail what happens daily in all state schools and, hence, influence very significantly the educational experience of every child: "We now have the benefit of a sound evidence base about *what needs to happen in every classroom and in every school…*" (Treasury, 2011b, p. 1) (emphasis ours). Moreover, the authors state "Our concern is to ensure that resources are directed to where they will have the greatest impact on student achievement. In our view, this is best done through a focus on ensuring effective teaching across the system" (Treasury, 2011b, p. 4).

Such claims and the broad policy initiatives which they support cannot go unchallenged by the research community. It is important that educationists query this policy agenda and critically examine the claims and assumptions which are meant to underpin it. This is what we attempt to do in the remaining sections of this paper. Having examined the assumptions that underpin Treasury's advice and the policy agenda it proposes to government, it seems clear that some or all of the premises on which the advice and agenda are based are, at the very least, questionable.

Public concern is warranted not least because the advice appears to propose quite specific education policy solutions without having demonstrated that: (1) there is an educational problem; (2) even if there is an educational problem it is a problem of teaching quality; (3) the proposed policy solutions have any practical prospect of improving the quality of teaching; and (4) that any improvements to the quality of teaching would lead to any improvements in student outcomes, let alone the magnitude of improvements claimed.

In contrast, our position is that the fact that a significant proportion of school students do not reach their full educational potential is considerably more complex than Treasury has allowed for. In our view, it is these complexities that must be better understood if we are to identify more precisely where changes in economic, social and educational policy settings are required to better support underachieving students. Consequently, it is these greater complexities which we explore in the remainder of our report.

SUMMARY

This section has shown that:

- Treasury's education policy advice is based on conventional economic theories of business process and human capital. Accordingly, it is assumed that variations in the quality of teachers and teaching can be removed over time through the use of incentives based on value added student achievement data, and that this in turn will resolve the problem of structural education inequalities.
- Treasury's analysis is simplistic and largely ignores or discounts other plausible contributory causes to educational inequalities. There clearly are important social and economic inequalities outside the control of the school, and some parents invest considerably more time and money than others in supporting their children's education.
- It is also assumed that value added assessment measures and performance related pay incentives will: (1) be valid, reliable and practicable; (2) be motivational to teachers; and (3) produce no unintended negative consequences. However, such assumptions are based on a limited evidence base of one particular economic kind. The assumptions also need to be subjected to critical educational scrutiny before such a radical departure from current policy settings is countenanced for New Zealand children and teachers. (See Sections 5 and 6)

SECTION 2. EFFECTIVE TEACHING: AN INTERNATIONAL PERSPECTIVE

> *The first and most solidly based finding is that the largest source of variation in student learning is attributable to differences in what students bring to school–their abilities and attitudes, and family and community background (OECD, 2005, p. 2).*

Several years ago the Organisation for Economic Cooperation and Development (OECD) conducted a study which claimed to reveal much about school teachers: their preparation, recruitment, work and careers (OECD, 2005). It involved 25 countries around the world. New Zealand was not included. The report began by making the point that:

> Student learning is influenced by many factors, including students' skills, expectations, motivation and behaviour; family resources, attitudes and support, peer group skills, attitudes and behaviour' school organisation, resources and climate; curriculum structure and content; and teacher skills, knowledge, attitude and practices. Schools and classrooms are complex, dynamic environments and identifying the effects of these varied factors, and how they influence and relate to each other – for different types of students and different types of learning – has been, and continues to be a major focus of educational research (OECD, 2005, p. 2).

The authors go on to say that three broad conclusions emerge from the research on student learning: (1) the largest source of variation in student learning is attributable to differences in students' abilities, attitudes and backgrounds; (2) within the school, teacher quality is the most important influencing variable; and (3) most research on teacher quality has been too narrow, focusing on students' test scores and teacher characteristics that are easily measured.

> The teacher characteristics that are harder to measure, but which can be vital to student learning include the ability to convey ideas in clear and convincing ways; to create effective learning environments for different types of students; to foster productive teacher-student relationships; to be enthusiastic and creative; and to work effectively with colleagues and parents (p. 2).

As we showed in Section 1, official proposals for New Zealand education currently being canvassed would embark on a programme of teacher assessment that ignores the clear qualifications set out by the OECD.

A basic and very important problem facing school systems can be set out in the following way.

1. According to the OECD report "teaching is a profession in long-term decline. As societies have become wealthier and educational qualifications have increased and employment opportunities expanded, teaching's appeal…does seem to have diminished" (p. 5).
2. However, there is increased awareness of how crucial teachers can be in the achievement of students and in the progress of society.
3. Yet many societies, including our own, seem bent on instituting measures (such as constant surveillance and performance management) which alienate teachers, deter students (especially the most gifted) from a teaching career, encourage unprofessional behaviour (such as cheating on tests) and lead to rapid turnover of teachers when the system needs stability.

These concerns will be developed and elaborated in the pages to follow. Our analysis suggests that the policy "solutions" currently being advocated may in fact exacerbate the problem.

Below we elaborate on the key points made by the OECD.

1. Social Class and Home Background[2]

The effects of socioeconomic status (SES) and home background on achievement are typically measured using a range of quantitative variables which allow researchers to identify the operation of relative deprivation or advantage across social groups. Researchers differ on which of these variables (individual, family, social, community) are most important, their modes of interaction and the consequent magnitude of their effects. Nevertheless, their

[2] This section draws on the work of two of us, Snook & O'Neill (2010).

studies all show that there is a consistent "gap" in average educational achievement between the children of high and low SES parents.

The "gap" is not restricted to one society (e.g. USA or NZ) or to one type of society (e.g. English-speaking); it occurs in every developed society. Students with good family resources typically out-perform those without (Biddulph, Biddulph & Biddulph, 2003). Gray, Jesson and Jones (1986) summarise their large scale research in Great Britain: "Around 80% of the difference can be explained by the intake" and they say that "this has held up over all the schools and LEAs studied" (p. 77). They go on to say that half the remaining difference (the 20%) may be explained by the school's examination policies. This would leave only 10% to be explained by other variables within the school.

Because of the socio-economic gaps, there is an "enormous disparity in children's home backgrounds and the social and cultural capital they bring to the educational table" (Benn & Millar, 2006, p. 23) And, despite its support for "accountability based programmes", the US Office of Education, having reviewed the international evidence, admitted that it was clear that "most participating countries do not differ significantly from the United States in terms of the strength of relationship between socioeconomic status and literacy in any subject" (Lemke, et al, 2002, p. 35).

When children attend schools which are widely different in social class composition, the gaps between the aggregated achievement of schools mirror closely the gaps between the social classes which predominate in them. Based on his research in New Zealand (and consistent with many overseas studies), Richard Harker has claimed that "anywhere between 70-80% of the between schools variance [see Appendix A] is due to the student 'mix' which means that only between 20% and 30% is attributable to the schools themselves" (Harker, 1995, p. 74).

As one part of a significant longitudinal study of over 1,000 NZ children born in 1977, researchers analysed the statistical relationship between family SES in childhood and educational advantage. The authors reported "the presence of pervasive relationships between SES at birth and material, cognitive, family and school factors" (Fergusson, Horwood & Boden, 2008, p. 20).

The OECD is correct to acknowledge that student ability, attitude and background are "difficult for policy makers to influence, at least in the short-run" (OECD, 2005, p. 2).

It is strange, then, that despite the overwhelming research findings relating to the effects of ability, attitude and background, politicians and the popular media persist in minimising their significance and insist that great improvement in learning can be gained by concentrating on schools and teachers in isolation from what occurs in students' lives outside school. There is little doubt that fundamental change in student achievement patterns would require coordinated efforts in social and community development, alongside efforts within schools and improved teacher education.

2. Teacher Quality

There is plenty of evidence that within formal educational settings, teaching practices in individual classrooms exert greater influence than do institutional level variables. Here research reinforces common sense. Nevertheless, even the claim about the centrality of the teacher as the major variable "within the school" has to be looked at very critically. For one thing, student ability and home background are not just variables at the point of entry: they remain operative throughout the process. Ability (however this is construed) and home environment continue to affect the learning of students at all stages ("the rich get richer and the poor get poorer"). The study by Fergusson and colleagues makes this clear: "family SES assessed at birth had multiple consequences for the economic, social, and educational environment within which the child was reared" (p. 24). Family SES affects not only the acquisition and use of resources but also the values, attitudes, behaviour and physical environment of individuals. These continue to make a difference to the educational progress of students and hence any assessment of that progress must acknowledge that there are many other influences on a student's learning besides the quality of the teacher. For these and other reasons, some statisticians argue that test scores can *never* be a reliable way of assessing the quality of teachers. Thus, after reviewing six major models of Value Added Measures (VAM), Sloane and colleagues (2013) concluded that "VAM –based estimates should never serve as a single indicator of teacher effectiveness, and high-stakes decisions should never be made primarily on the basis of VAM-based estimates" (p. 64). (See Sections 5 and 6)

3. Narrow Research on Teaching

In stating that much research on teacher quality has focused too heavily on test scores and "readily measurable" teacher characteristics, it is interesting to see that the OECD is echoing what teachers and educational researchers have been saying for quite some time, while politicians and administrators have forged ahead with proposals to "measure" student learning and teacher's competence by focusing narrowly on test scores. As we shall see, today's politicians and administrators are still inclined to recommend and legislate such narrow forms of accountability. In contrast, we argue that teaching is much too complex for such narrow evaluation to be meaningful or fair. Student learning is a function of all of the following variables.

The skills, attitudes and values which the learner brings to the classroom;

1. The developmental appropriateness of the learning activities provided;
2. The number of learning interactions experienced by the learner, their types, and the way in which they are sequenced;
3. The appropriateness of the teaching and practice activities provided given the type of learning aimed for;
4. How student improvement is monitored over time;
5. How the results of this monitoring are used to adapt teaching and practice activities for each student from day to day; and
6. School and classroom contexts including the contributions made by other students.

As we shall see later (Section 4) all of this complexity must be recognised and coped with if we are to devise ways of accurately assessing the effectiveness of teachers. This is especially so if the assessments are high stakes (i.e. involving performance pay, promotion, suspension or dismissal). (See Sections 5, 6, 7)

SUMMARY

This section has shown that:

- The socio-economic background and prior life experiences of students *outside school* are the main influences on learning. Their effects need to be fully taken into account if they are not to be overlooked in the creation of school-based "solutions" which cannot work and which merely alienate teachers when their morale and cooperation are most required.
- Despite the OECD's claim (correct as far as it goes) of the critical importance of teachers *within the school*, it is important to remember that, in another way, the student is of most importance: what she brings to the classroom (natural ability, cultural norms, values, knowledge, skills) plays a central role in what and how she learns.
- The OECD, having studied twenty five educational systems warns against trying to measure teacher effectiveness on the basis of tests administered to students and on those characteristics of teachers which, though easily measured, have not been shown to be reliable measures of "effective teaching".

SECTION 3. INTERNATIONAL STUDIES OF STUDENT ACHIEVEMENT

> *Nations with high-stakes testing have generally gone down in scores from 2000 to 2003, and then again by 2006. Finland, on the other hand, which has no high-stakes testing, and an accountability system that relies on teacher judgment and school level professionalism much more than tests, has shown growth over these three PISA administrations (Sahlberg, 2011).*

The 2005 Report of the Educational Review Office (ERO) stated that: "New Zealand's best students perform with the best in other countries but there is a group at the bottom, perhaps as large as 20 per cent, who are currently not succeeding in our education system" (ERO, 2005, p. 6).

This led the Education and Science Committee to set up the *Inquiry into Making the Schooling System Work for Every Child* (2008). Since then "the long tail of underachievement", as it has come to be called, has received much attention in the media and in official documents (e.g. Treasury, 2009, 2012). For example, the Briefing to the Incoming Minister of Finance (BIM) from Treasury following the 2011 election stated that:

> New Zealand has a wide distribution of educational achievement and more low performing students compared to other countries with a similarly high average score in international tests (OECD, 2010a). As a result of this wide distribution, a McKinsey and Company report on high performing schooling systems classifies New Zealand's schooling system as 'fair to good', similar to countries such as Malaysia, Armenia and Portugal (Mourshed et al., 2010) (Treasury, 2012, p. 1).

Within this statement are two claims that warrant further investigation:

1. That New Zealand has a wider distribution of student achievement than similarly achieving countries; in other words "The long tail of underachievement".
2. That due to this wide distribution of achievement, New Zealand's education system is considered only "fair to good" by the McKinsey Report.

We will examine each of these claims separately.

1. The Long Tail of Underachievement

The claim is that New Zealand has a wider distribution and more low performing students than countries with similar average scores. The evidence cited by Treasury to support this claim comes from the 2009 Programme for International Student Assessment (PISA) an international study that assesses a random sample of 15-year old students' achievement levels in the areas of reading literacy, scientific literacy and mathematics literacy. PISA assesses students from 65 countries or economies, including both OECD and non-OECD jurisdictions. The PISA study is conducted every three years and all three content areas are assessed with a particular focus on one specific area in each cycle. The last available results were for the 2009 round that focused on reading literacy.

Data from the 2009 Programme for International Student Assessment (PISA)

New Zealand has traditionally performed well in PISA. The average scores in reading, science, and mathematics literacy have been significantly above the OECD average in all three areas since the inception of the study in 2000. In the 2009 round New Zealand was 7th in reading literacy, 7th in scientific literacy, and 13th in mathematics literacy. Across all three disciplines, New Zealand's mean performance is statistically similar to that of: Canada, Japan, and Australia. Only Finland, and two non-OECD education systems, Shanghai-China and Hong Kong-China, consistently achieved an average score higher than New Zealand. Significantly, the majority of OECD countries, including the UK and the USA, perform much lower than New Zealand (Ministry of Education, 2010a).

The Treasury and other commentators on the achievement of New Zealand students have acknowledged the evidence that New Zealand students performed very well overall in the 2009 PISA assessments, but have expressed concern about the wide spread of achievement levels for 15-year-old New Zealanders and the percentage of students performing poorly. Here, we look at how accurate and fair those comments have been. PISA has classified achievement into seven bands of performance in reading and into six bands in mathematical and scientific literacy. In the section about each performance area which follows, we report the percentages of New Zealand students who scored in low and high achievement bands. To allow comparisons, we also present the average percentages for all OECD countries and the percentages for six other OECD countries: Australia, the UK, the USA, Canada, Japan and Finland (the top or second top OECD country in all three assessments).

Table 1: Percentages of students in selected countries scoring low or high in PISA Reading Literacy 2009 (Note: seven bands)

	Per cent in Bottom 2 Bands	Per cent in Bottom 3 Bands	Per cent in Top Two Bands	Per cent in Top Band
New Zealand	4	14	16	3
Australia	4	14	13	2
UK	5	18	8	1
USA	4	17	10	1
Canada[3]	2	10	13	2
Japan	5	13	14	2
Finland	1	8	15	1
Average for all 34 OECD Countries	6	18	8	1

New Zealand had the highest percentage of students scoring in the top performance band, higher than in any of the other 33 OECD countries (Table 1). The same is true for performance in the top two bands. At the other end of the performance scale, only two of the comparison countries (Canada and Finland) had a lower percentage of students than New Zealand scoring in the bottom two bands, while Finland, Canada and Japan had lower percentages of students in the bottom 3 performance bands. While it is true that New Zealand has a wide range of achievement in reading literacy, that arises because New Zealand has the largest proportion of high achievers in the OECD, not because it has a particularly high proportion of low achievers. Only three of the 34 OECD countries (Korea, Finland and Canada) had markedly lower percentages of students in the bottom three performance levels for reading literacy, with Japan and Estonia just 1% lower than New Zealand. Twenty-eight OECD counties had higher percentages of students in the bottom three performance levels. While there is always scope for improvement in our students' achievement, it clearly is not true that New Zealand overall has a serious issue of low achievement in reading literacy.

[3] Canada excludes First Nation People from its statistics. This gives it a distinct advantage over other countries (such as New Zealand) which include the statistics for indigenous people.

Table 2: Percentages of students in selected countries scoring low or high in PISA Mathematical Literacy 2009 (Note: six bands)

	Per cent in Bottom Band	Per cent in Bottom 2 Bands	Per cent in Top Two Bands	Per cent in Top Band
New Zealand	5	15	19	5
Australia	5	16	18	4
UK	6	20	10	2
USA	8	23	10	2
Canada[4]	3	12	18	4
Japan	4	12	21	6
Finland	1	8	22	5
Average for all 34 OECD Countries	8	18	13	3

Only one of the six comparison countries, Finland, had a substantially lower percentage of students scoring in the bottom two performance bands in mathematical literacy, and none of the comparison countries was substantially ahead of New Zealand in the percentage of students in the top two performance bands (Table 2). Considering the data for all OECD countries, only four countries (Korea, Switzerland, Japan and Belgium) had higher percentages than New Zealand of students scoring in the top performance band. Those same four countries plus Finland and the Netherlands had higher percentages of students in the top two bands. At the other end of the performance scale, ten OECD countries had lower percentages than New Zealand of students scoring in the bottom band, and seven OECD countries had lower percentages of students in the bottom two bands. Only Finland and Korea had less than 12% in the bottom two bands. Thus in mathematical literacy, as in reading literacy, New Zealand has a slightly wider range of achievement than many other countries, but this arises as much from good performances by our higher scoring students as from poor performances by our lower scoring students.

[4] See Footnote 3

Table 3: Percentages of students in selected countries scoring low or high in PISA Scientific Literacy 2009 (Note: six bands)

	Per cent in Bottom Band	Per cent in Bottom 2 Bands	Per cent in Top Two Bands	Per cent in Top Band
New Zealand	4	13	18	4
Australia	3	12	15	3
UK	4	15	12	2
USA	4	18	10	2
Canada[5]	2	9	13	2
Japan	3	10	17	3
Finland	1	6	19	4
Average for all 34 OECD Countries	5	18	9	2

Two of the six comparison countries, Finland and Canada, had substantially lower percentages of students scoring in the bottom two performance bands than did New Zealand, and none of the comparison countries was substantially ahead of New Zealand in the percentage of students in the top two performance bands (Table 3). Considering the data for all OECD countries, no country had a higher percentage than New Zealand of students scoring in the top performance band, and only Finland had a higher percentage than New Zealand scoring in the top two bands. At the other end of the performance scale, eleven OECD countries had lower percentages than New Zealand of students scoring in the bottom band, and ten OECD countries had lower percentages of students in the bottom two bands. Finland, Korea, Estonia and Canada had less than 10% in the bottom two bands. Thus again, in scientific literacy, New Zealand has a slightly wider range of achievement than many other countries, but this arises as much from good performances by our higher scoring students as from poor performances by our lower scoring students.

Some issues associated with the 2009 PISA assessments

The 2009 assessments in all three areas present a generally positive picture of overall achievement of New Zealand 15 years old students in reading literacy, mathematical literacy

[5] See Footnote 3

and scientific literacy. Very few OECD countries scored higher, in any of the three areas. This arose from comparatively high proportions of students scoring highly, and moderate proportions of students with low scores.

New Zealand boys and girls performed very similarly in mathematical literacy and scientific literacy. As was the case in other countries, New Zealand girls scored higher than boys in reading literacy, but the gap was a little larger in New Zealand than in most other countries. Compared to the overall (boys plus girls) results for the other 33 OECD countries, New Zealand girls scored higher than all of the countries, while 13 countries scored higher than New Zealand boys.

Students of all ethnicities are included amongst New Zealand's very highest performers and among the very lowest performers. Nevertheless, it has been apparent for a long time that higher proportions of Pasifika and Māori students than Asian and New Zealand European students are among our low scoring students. As a group, New Zealand European students scored as well or better than the overall results for the other 33 OECD countries in both reading literacy and scientific literacy and just below the top two countries in mathematical literacy. In all three areas, Pasifika students as a group scored higher than the students in just one country (Mexico) and at a similar level to the students in only two countries (Chile and Turkey). Māori students as a group scored a little higher than Pasifika students in each of the three areas. These patterns reaffirm that efforts to better address the learning needs of Pasifika and Māori students will be crucial as we try to reduce New Zealand's percentage of low performing students.

Overall, these results do not support the Treasury's claim that New Zealand has a "long tail of underachievement" in relation to similarly achieving countries. Rather they show that on PISA New Zealand has both a higher than average proportion of students in the higher proficiency levels and a lower than average proportion in the lower levels of reading, mathematical, and science literacy. This indicates that, although New Zealand might have a wide spread of achievement, this spread is situated towards the upper end of the achievement spectrum (Ministry of Education, 2010a). High achieving students are doing better and lower achieving students are generally performing on a par with countries that are statistically similar. Although this might suggest that we could do better, it certainly does not suggest a poorly performing school system.

Data from TIMSS and PIRLS

Although the Treasury cites PISA to support its claim that New Zealand has a wider distribution and more low performing students than countries with similar average scores, there are other internationally recognized measures of student achievement which merit attention in relation to New Zealand's international standing. Two of these are the Trends in International Mathematics and Science Study (TIMSS), which assesses students' achievement in mathematics and science in Y5 and Y9, and the Progress in International Reading Literacy Study (PIRLS), which assesses reading literacy at Y5. TIMSS is conducted on a four yearly cycle and PIRLS on a five yearly cycle. Approximately 60 countries including 28 OECD countries took part in either TIMSS or PIRLS in the most recent assessments, conducted in 2010/11.

The TIMSS results reveal some differences from the PISA results. In the 2010/2011 round, relative to other participating OECD countries in particular, New Zealand students performed much more modestly in mathematics and science than our students had in the 2009 PISA round, hovering around the TIMSS scale centre-point. The Y5 mathematics result was statistically just below the centre-point, Y9 maths and Y5 science were statistically similar to, and the Y9 science was statistically just above, the scale centre-point. At Y5, relative to all countries we may be said to perform "poorly" in all mathematics and science; relative to OECD countries, we do "very poorly". At Y9, relative to all countries in mathematics we perform "well", but "poorly" relative to OECD countries. In science, relative to all countries we perform "extremely well"; relative to OECD countries, we perform slightly better than average. The range of scores between New Zealand's highest and lowest performing students was considered similar to most other countries in both subjects and at both age levels (Ministry of Education, 2013).

The PIRLS data (reading achievement in Yr 5) also shows some differences from the PISA data. Although New Zealand (531) scored statistically much higher than the scale centre point of 500, our median was towards the lower end of the OECD countries. Additionally, our range of scores was relatively large when compared with other countries where English was one of the assessment languages (Ministry of Education, 2012).

There are four international benchmarks used in TIMSS and PIRLS to separate students' achievement into levels: "advanced", "high", "intermediate" and "low". At Y5 mathematics

and science, the proportions of New Zealand students reaching each of the benchmarks were generally not as high as the international median proportion reaching those benchmarks. In contrast, at Y9, the proportions of New Zealand students reaching the benchmarks in both subjects were higher than the international median proportion of students reaching those benchmarks (see Table 4). Furthermore, while the proportion of Y5 mathematics and science students reaching the advanced level is the same as the international median, at Y9 there are about twice as many reaching this level as the international median. Put simply, younger students in New Zealand are not performing as well in maths and science at Y5 but they catch up and surpass the international median levels by the time they reach Y9. PIRLS tests only at the Y5 level and again we see that the proportion of students not reaching the low and intermediate international benchmark is higher than the international median proportion at this level. However, according to PISA (the only available test of reading at a later age) reading achievement has greatly improved by age 15.

Table 4: Percentage of students not reaching low international benchmark (PIRLS, TIMMS 2010/2011)

	Y5		Y9	
	NZ	International median	NZ	International median
Reading	8	5	-	-
Mathematics	15	10	16	25
Science	14	8	10	21

Commenting on the "underperformance" of New Zealand Y5 students in the TIMSS and PIRLS the Ministry of Education (2012) argues that:

> … there are a significant number of schools that are presented with challenges associated with the composition of their student body being from predominantly economically disadvantaged backgrounds and their student intake not having the necessary early literacy or numeracy skills, or language skills when beginning school. (p. 33)

In other words, a student's socioeconomic background has a significant impact on their achievement and this is visible on entry to school. Māori and Pasifika students are over

represented in the lower socioeconomic levels in New Zealand. They are also over represented at the lower levels of achievement in TIMSS and PIRLS.

PISA and social equality

Among the OECD countries, in the late 2000s, New Zealand was one of the most unequal countries in terms of income distribution, ranking 28th out of 31 countries on the Gini Index. Between the mid 1980s and mid 2000s, income inequality rose in 15 of the 19 OECD countries for which data are available. New Zealand (together with Sweden) had the highest cumulative rate of change in this period. [6] In the same period, there were significant increases in the numbers of people living in poverty (below 50% of median income) in New Zealand (OECD, 2008, Table 11.1). In the mid 2000s, 15% of children were living in poverty, compared with the OECD average of 12%. The rate of increase in child poverty since the mid 1990s was 2.3%, compared with the OECD average of 1.0% (OECD, 2008, Table 5.2).

One measure of the impact of socio-economic status is the slope of the graph of the PISA mean score against the index of economic, social and cultural status (ESCS). The steeper the gradient, the greater the apparent impact of economic, social, and cultural status on students' performance. The OECD average for this gradient is 38, indicating that a one point reduction in the ESCS index would raise a student's score by 38 points, roughly equal to one year's worth of schooling. The New Zealand value of 52 is the highest value in the OECD countries (OECD, 2010a). That is to say, the impact of economic, social and cultural status on student's performance in New Zealand is the highest among the OECD countries.

A second measure of the impact of socio-economic status on performance is the measure of the "strength of the gradient". This is a measure of the percentage of variance in students' performance explained by their socio-economic background. The larger the value, the more the variation in a students' performance can be attributed to their socio-economic background, and the more accurately their socio-economic and cultural status can predict their achievement (OECD, 2010a). The value for New Zealand is 16.6, higher than the OECD average of 14.0, and significantly higher than the values for Australia (12.7), Canada (8.6) and Japan (8.6): countries that have a statistically similar mean PISA score to New

[6] http://www.oecd-ilibrary.org/sites/factbook-2013-en/

Zealand. Put simply, the likelihood of disadvantaged students performing at levels similar to those of their advantaged peers is lower in New Zealand than in countries such as Australia, Canada, and Japan.

Despite this, the proportion of "resilient" students—those who come from a disadvantaged socio-economic background and perform much higher than would be predicted by their background—is 9.2% of all New Zealand students, statistically higher than the OECD average of 7.7% (OECD, 2010a). Moreover, if we take this as a proportion of those who are considered disadvantaged then approximately 36% of disadvantaged students can be considered "resilient", higher than the OECD average of 31% (OECD, 2010).

The resilience data indicate that New Zealand teachers have been as effective or more effective than the teachers in comparable countries in helping to overcome the effects of poverty and low SES on school achievement. This data provides no support whatsoever for Treasury's claim, repeated by the Minister of Education and the previous Director of Education, that New Zealand teachers have been less successful than others in overcoming social disadvantage.

For a further investigation of equity in the school sector based on socioeconomic grounds or ethnicity we can examine the trend data from New Zealand's National Education Monitoring Programme (NEMP).

Findings from NEMP

The National Educational Monitoring Programme (NEMP) was designed to get a broad picture of the achievements of students in New Zealand schools. It ran from 1995 to 2010 and assessed representative samples of students at Y4 and Y8 in 15 curriculum areas over a four year cycle. Rather than providing comparisons with other countries, this New Zealand based project was designed to identify national trends in educational performance to assist policy makers, curriculum specialists and educators with their planning. One aspect of the NEMP reports that is useful for this discussion is the trends in students' achievement levels between schools of different decile levels and also between different ethnicities.

Looking initially at the decile level data, we see that students from higher decile schools outperform mid decile schools, which in turn outperform low decile schools. This pattern has

applied typically for 50% to 80% of the assessment tasks in each curriculum area assessed, with no consistent trend downward or upward across 12 year periods of assessment. The exceptions to this pattern have occurred for the assessments in Physical Education and Music, where differences have been found generally on less than 50% of tasks (usually 10% to 30% in physical education and 30% to 50% in music), but again no consistent upward or downward trend has been observed (Ministry of Education, 2007, 2008, 2009, 2010b, 2011).

The ethnicity data compare the achievement levels between Pakeha and Māori, and between Pakeha and Pasifika students. Pakeha students generally perform better than Māori students (effect sizes average about 0.3, and with one exception range between approximately 0.2 and 0.4 – the exception is for physical education, in which Māori student have performed slightly better than Pakeha students). Similarly, Pakeha students generally perform better than Pasifika students (effect sizes average about 0.4 and with two exceptions range between approximately 0.3 and 0.6 – the exceptions are for physical education and music, with Pasifika students perfoming slightly better than Pakeha students in Physical Education and only slightly lower in Music). There have been fluctuations in disparity for individual subjects, but overall there has been little evidence of any trend over the last eight years of assessments (Ministry of Education, 2007, 2008, 2009, 2010b, 2011).

Overall, the trend data reveal no significant movement in reducing the disparities that exist between schools of different deciles or between students from different ethnicities. Given the over-representation of Māori and Pasifika students in the lower socioeconomic levels in New Zealand, it is hard to separate out those aspects of poor performance that relate to ethnicity and those that relate to poverty. In both instances, decisions about how to address such issues of disparity must be broad national policy decisions because the disparities cannot be ameliorated by educational policy alone. In any event, the responsibility to address such issues of disparity does not rest realistically with individual teachers or schools.

When we look at findings for particular areas of the curriculum, it appears that there have been:

- Slight improvements from 2002 to 2006 in writing convention but these are too small to be significant; slight improvements in functional writing for Year 4 and only slight (but not significant) increases in Year 8: substantial increases in expressive writing in Year 4 but only modest increases in Year 8.

- Little change overall in science from 1995 to 2007; no significant changes in "the living world and planet earth"; a small but noteworthy decline in "the physical world" and a noteworthy decline in "the material world".
- Over the years there has been a small but not significant improvement in oral reading, no change in oral description and oral presentation and a marginal decline in reading comprehension. Since 1995 there has been a substantial improvement in reading and speaking in Year 4 and a small but significant improvement in Year 8.
- Since 1997 there has been a small improvement in mathematics in Year 4 but no meaningful improvement in Year 8. Since 2005 there has been no significant change in any of the areas (number, algebra, measurement, geometry, statistics). There was a slight decline in quick recall and the derivation of number facts.

On balance it would seem that while there has been no decline in student achievement at either level, there has not been any significant improvement either.

The Treasury's claim

We must now return to the Treasury's claim that New Zealand has a wide distribution of student achievement and more low-performing students compared to other countries with a similarly high average score in international tests. Although, the PISA data do show that we have a wider range than similarly achieving countries/ economies, the wide distribution can be predominantly explained by our top students doing well: the Treasury's claim ignores the fact that our distribution of student achievement is situated at the higher end of the achievement scale. In essence, our range starts at a similar level but extends higher than most similarly achieving countries. Furthermore, the data do not support the claim that New Zealand has more low-performing students compared to similarly achieving countries. Accordingly, in their summary report for PISA 2009, the Ministry of Education (2010a) concedes that:

> Although New Zealand continues to show a wide range of scores in reading, not all this difference can be attributed to low performing students. The success of the highest performing students also increased the size of the spread (p.14).

The TIMSS data tell us that, compared to other similar countries/economies, performance in mathematics and science is not impressive at Year 5. However, our students improve substantially as they get older, both in their average performance, and in the proportion of students meeting higher international benchmarks. The PIRLS data tell us that there is a relatively wide spread of achievement in reading with higher proportions of lower achieving students at Y5 than the international median (See Table 4). The only other reading assessment is PISA at age 15, which shows our lower end performance to be similar to countries/economies comparable to us and also shows particularly large proportions of our students reaching higher benchmarks.

Caution

While these studies provide us with some insights into the achievement levels of New Zealand students, they are a partial picture. As such, we should be cautious about over-reacting to some of the messages within them. Different assessment tools will produce different measures of achievement based on the particular subset of skills and knowledge they assess. PISA is aimed at testing the application of knowledge and problem solving: TIMSS is more focused on traditional classroom content and curriculum knowledge (OECD, 2010b). As such, it is likely that these two international studies will produce different achievement patterns depending upon the policy focus within a given country. Additionally, although the studies might identify areas of content or skills where students are relatively stronger or weaker nationally, this could be related to the structure and content of the curriculum (for example, if we are weaker in algebra in Y9 this may be because students generally do not start their study of algebra until Y9) or preferred pedagogy (for example, officially endorsed methods for teaching children how to read), or a shortage of teachers with the requisite specialist curriculum knowledge and pedagogy (Education Review Office, 2012). In many instances these are national policy choices and mandates rather than teacher or school decisions. Over the longer term, such policy "settings" have practical consequences, for example on the time spent on a particular learning area, and the range and quality of learning activities undertaken.

> TIMSS reports that in 2006 Year 5 students in New Zealand spent an average of 45 hours a year on science (down from 66 hours in 2002) and only six participating countries reported spending less time on science…. NEMP data show that in 2007 more students at both Year 4 and Year 8 indicated that their class 'never' did

experiments with everyday things, experiments with science equipment, or visited science activities than in 1999 (Gluckman, 2011, p. A-32).

2. The New Zealand System is Rated "Fair to Good"

The Treasury's second claim is that, because of our wide distribution of scores, the NZ educational system has been rated only "fair to good" by the third McKinsey report. A closer inspection of the report reveals that although it investigated 20 educational systems from around the world in some depth, *New Zealand was not among them*. Indeed, New Zealand is mentioned only twice in the report, once on a graph depicting various countries achievement levels in international assessments (PISA, TIMSS, PIRLS etc.) against public expenditure per student, and once in the list of OECD countries that acted as a control group in developing a "Universal Scale".

To enable comparisons across the various international measures of achievement, the report uses a "Universal Scale" score based on the methodology developed by Hanushek and colleagues (OECD, 2010). This methodology "normalizes" a country's score for each of the various international measures of student achievement, providing an average score for that country on a common scale. It does this by focusing on the transformations of the means and variances of the set of original country scores for each of the various international measures of student achievement. For example in providing a measure of New Zealand's performance on PISA 2009, the methodology considers the distribution of the mean scores of the 65 countries that took part that year. It then looks at the mean and the variance of that distribution to conduct the transformation to the Universal Scale score. It does not consider the individual variance of any given country's distribution of achievement scores.

As such, even if we accept New Zealand's position on this Universal Scale, the claim that it is our wide distribution of achievement scores that has placed us in a particular position on the scale is not supported by the methodology used to create it.

We submit therefore that all of the Treasury's claims about New Zealand's performance in international assessments are false: (1) There is no "long tail of underachievement" in any general sense; (2) New Zealand teachers are not deficient in remedying socio-economic disadvantage; and (3) The claim that because of the "long tail" of underachievement, New

Zealand education is only “fair to good” is unsubstantiated by the Report which is used to justify the claim.

SUMMARY

This section has shown that:

- There are several international assessments undertaken periodically to compare the academic performance of New Zealand school students to that of their overseas counterparts. New Zealand students generally perform well in these assessments, and particularly well by the middle of secondary school (PISA assessments). However, politicians and officials have asserted that there is a significant minority of students who do not perform well, hence the term "long tail of underachievement". While it is true that some New Zealand students do not achieve at school, both the size of the so-called tail and the reasons for it are hotly debated.
- Overall, the results do not support the Treasury's claim that New Zealand has a "long tail of underachievement" in relation to similarly achieving countries. For example, on PISA, New Zealand has both a higher than average proportion of students in the higher proficiency levels and a lower than average proportion in the lower levels of reading, mathematics, and science literacy. This indicates that, although New Zealand might have a *wide spread of achievement*, this spread is situated towards the upper end of the achievement spectrum. High achieving students are doing better and lower achieving students are generally performing on a par with countries that we like to compare ourselves with.
- International studies (such as PISA) are not studies of teacher quality. They are studies of system performance. Although these studies provide interesting data regarding the achievement of New Zealand students when compared against that of students in other countries, student achievement is a function of many factors in addition to teacher effectiveness. Student achievement is also affected by the curriculum (what teachers are required to teach), by policies regarding how teachers are to teach (how reading is to be taught, for example) and by funding allocations (for teacher training, for teaching resources, for salary levels, for staffing levels and so on). Accordingly, in order to provide a balanced picture, these studies should be considered in context, along with other national measure of student achievement, in light of the international research literature, and in consultation with expert educational groups.

SECTION 4. ASSESSING TEACHER EFFECTIVENESS

> *The result of studies of 'best' teachers is usually a picture of what the experts currently deem best. Whatever is fashionable at the time determines what researchers look for and what they see (Nuthall 2007, p. 29).*

The definition of effective teaching has changed over time and methods for assessing teachers have changed as beliefs about what is important to measure have evolved (Goe, Bell & Little, 2008). Although there is a general consensus that good teaching matters and may be the single most important *school-based* factor in improving student achievement, development of a reliable measure of teacher effectiveness has remained elusive (Goe, et al., 2008). Indeed, it has long been recognized that "an enormous underlying problem with teacher evaluation relates to lack of agreement about what constitutes good or effective teaching" (Cruickshank & Haefele, 1990, p. 34).

Goe et al. (2008) claim that effective teachers are those who:

- Set high expectations for every student with every student (including those with special needs) receiving the help needed to meet these expectations;
- Ensure that classroom life fosters academic and social outcomes such as increasing self-regulation and self-efficacy, positive attitudes to learning, cooperation and other essential social skills;
- Monitor each child's progress and adapt instruction to the needs of individual learners;
- Collaborate effectively with other teachers, administrators, parents and educational professionals;
- Contribute to the maintenance of a positive and well ordered school.

The assessment of teachers has always been a central concern and, from the beginning of our system in 1877, inspectors monitored teachers' behaviour in classrooms. Since the reforms of Tomorrow's Schools, the Education Review Office has reported on the performance of schools. In recent years, however, there have been demands for a more "formal" and "objective" evaluation of teachers (not just schools) and, often, for financial rewards for teachers in accordance with the results. It is clear that an evaluation of teaching must focus

on: (1) what the teacher does and says; as well as (2) what the student learns. Over the past half century, research on assessing teacher effectiveness has focused on these two areas: teacher characteristics/behaviour and student achievement.

1. Research on Teacher Characteristics

The earliest attempts to measure teacher effectiveness involved examinations of the relationship between teacher characteristics (as reported by principals) and student achievement. Most of these studies were undertaken in the first half of the twentieth century and by the 1970s most reviewers had concluded that the small and highly variable correlations which were being observed between ratings of teacher characteristics and student learning were too transient and unreliable to be used in evaluating teacher effectiveness: "Each such study has concluded that there is no appreciable agreement between principals' judgments of teachers' effectiveness and the amount students learn" (Medley & Coker, 1987, p.140) Indeed, "it has been known since the 1930's that ratings made by school principals and teacher educators do not correlate with students' achievement" (Nuthall, 2002, p. 47).

Research on teacher classroom behaviour

There have been numerous attempts to develop classroom observation protocols which can be used by senior school personnel or outside evaluators to assess teacher effectiveness. These have been reviewed by Goe and colleagues (2008) who found that there is little research into the use of classroom observations as a method for evaluating teachers' effectiveness in producing student learning. They also concluded that these kinds of measures "are highly dependent on the instrument used, sampling procedures, and training of raters" (p. 16). Nor can student learning be inferred from such observations. For example, one of the factors on many observation schedules is level of student engagement. But student engagement, while necessary for learning, is not learning. Students can be engaged in a classroom activity without learning anything from that activity.

2. Research on Student Achievement

Faced with these difficulties, it has increasingly been argued that teacher effectiveness should be measured by measuring the achievement of their students. The problem with this idea is that student achievement is a function of many factors: prior learning, peer support for learning, parental support for learning (and so on) as well as the efforts of the classroom teacher. Research into the many factors on which student achievement depends is extensive and includes research undertaken at both the primary school level (e.g. Cervini, 2008) and the secondary school level (e.g. Harker & Tymms, 2004; Palardy, 2007). These studies have examined the relationships between student achievement and home factors (such as parental education and income), neighbourhood factors (such as level of affluence and community support), student factors (such as ability and prior achievement), school factors (such as school mix, organisation and resources), and classroom factors (such as student mix and teacher effectiveness) (Teddlie & Reynolds, 2001).

Typical of this research are the findings of Opdenakker and van Damme (2000), in their study of the correlations between gains in mathematics achievement and student level, class level, teacher level and school level factors in Belgium secondary schools: “analysis of the four-level null model revealed that each level is important: 54.43% of the overall variance in mathematics is linked to the individual level, 14.71% to the class level, 10.07% to the teacher level and 12.81% to the school level” (p. 108).

Schochet and Chiang (2010) say: “our results are largely driven by findings from the literature and new analyses that more than 90% of the variation in student gains scores is due to student-level factors which are not under the control of the teacher” (p. 25). This is not surprising because children spend only a small part of their waking hours in classrooms. What children learn during the remainder of the time varies widely from one student to the next as a function of parental availability, education, income, aspirations, community resources and much more. Furthermore, the rate at which a child develops new skills and understandings is directly related to the number of relevant learning opportunities which he or she experiences from minute to minute, hour to hour and day to day. For example, the rate of vocabulary growth during the first three years is directly related to the number of child-adult conversational interactions that the child experiences from day to day (Hart & Risley, 1995).

Interactive effects

Students in classes learn not only from their teachers: they learn also from constant interaction with their peers and, most importantly, they learn by reconstructing their experience in the light of their own understandings, past knowledge and motivation. These two features of classroom life were highlighted by Graham Nuthall during decades of close observation in New Zealand primary classrooms. He found that prior knowledge differs dramatically from one student to the next. Because of these individual differences in prior knowledge, as well as differences in the way students engage with classroom activities and with each other, each student experiences the classroom differently. Nuthall's work also reveals that previous attempts to produce generalizable results failed to explain teaching and learning because they overlooked the interactive and contextual nature of teaching and, in particular, neglected the fact that "the way students behave and experience classroom activities is not only a direct function of teacher-managed activities but also a function of their on-going relationships with other students and of their own beliefs and previous experiences" (Nuthall, 2002, p. 46).

Classroom effects

Classes of children vary with respect to the proportion of children with behaviour problems and learning disabilities, the proportion of children from under-resourced families, the proportion of very able children, and so on. Even after the effects of prior achievement have been taken into account, gains in achievement depend in part upon class composition (Kyriakides & Creemers, 2008; Peetsma, van der Veen, Koopman & Schooten, 2005). Measures of class composition typically account for 10% to 15% of the variation in student achievement gains.

School effects

Schools also vary with respect to their available resources, their student intake, and the quality of their principal and teachers. Differences between schools typically account for 12% to 15% of the variation in gains in student achievement (Teddlie, Reynolds & Sammons, 2000).

It must be noted, of course, that the correlations between achievement gains and the various factors listed depend upon the accuracy with which achievement gains have been measured.

Unreliable measures of achievement gains result in unreliable estimates of the strength of the correlation between achievement gains and each of the various factors under examination (Harker & Tymms, 2004).

SUMMARY

This section has shown that:

- Studies of teacher effectiveness focus on what the teacher does and/or on what the student learns. All of the studies reviewed in this section have limitations and therefore suggest caution to policymakers who wish to link periodic assessments of teacher effectiveness with teacher registration, employment, career progression and remuneration.
- Almost all studies suggest that the contribution of the teacher to variation in student achievement is relatively small (10%-20%). Other influences are student background, school factors and class composition.
- These studies generally have been limited to measures of improvement in reading or mathematics, and have largely ignored all other schooling outcomes such as improvements in other curriculum areas, improvements in social skills, improvements in self-directed ability to learn and improvements in the unique talents of individual children.
- All of these studies leave a major part of the variation in student achievement unexplained. This is because most of the measures of achievement used are not very accurate and because some of the factors affecting achievement (e.g. home background) are extremely difficult to measure.

SECTION 5. VALUE ADDED MEASURES

> *Our results are largely driven by findings from the literature and new analyses that more than 90 per cent of the variation in student gain scores is due to the variation in student-level factors that are not under the control of the teacher* (Schochet & Chiang, 2010, p. 35).

Recognizing that teachers cannot be held accountable for effects outside their control (e.g. home background, previous knowledge, classroom composition, natural ability and mental engagement), researchers have turned to trying to assess what "value" a teacher has "added" to the student's learning over the time they were in her classroom. Thus has emerged a particular form of achievement based assessment known as Value Added Measurement (VAM).

In the basic version of this approach, researchers take two measures of student achievement (e.g. at the beginning of the year and at the end) and calculate the degree of improvement between those two measures. This improvement is attributed to the child's teacher in that year and thus the calculation is supposed to reveal which teachers are good and which are bad. This approach is seen as fairer than simply judging the teacher on results without regard for the different starting points of individual students. The basic model has been subjected to many kinds of qualifications in the hope of making the instrument more reliable. As we shall see when examining the English CVA (See Section 6 Case Study 2), it is doubtful whether even the most elaborate models are any more reliable than the basic one. Some of the major problems with most VAMs are as follows.

1. They take no account of differential student abilities. Ability and background are not just variables at point of entry: they continue to affect learning throughout the process. Ability (however this is construed) continues to affect the learning of students at all stages and the same is true of home background ("the rich get rich and the poor get poorer"). Both continue to make a difference to the achievement of students and hence any assessment of student progress must acknowledge that there are many other influences on a student's learning during a year of teaching. Gene Glass argues that all value added measures are faulty because:

> ... they act as though the statistical equalling on achievement tests (as fallible as it is) of groups of students has held all influences constant (ceteris paribus) and hence the gain score is valid and a fair as a measure of the contribution to learning of a teacher or a school. It is not and never will be (Glass, 2012, unpaginated).

For these and other reasons, some statisticians argue that test scores can never be a reliable way of assessing the quality of teachers. For example, after careful analyses of teacher assessment scores, a report from the National Center for Educational Evaluation concluded that, if three years of data were used, the result would be that more than 1 in 4 teachers who are truly average in performance would be erroneously identified for special treatment. If ten years of data are used, the error rate halves but is still significant (Schochet & Chiang, 2010). It is important to recognise what this means. Suppose that, as some have suggested, the lower achieving 20 per cent of teachers were to be dismissed. This means that if there were 100,000 teachers in a school district and three years of data were used, at least 5,000 teachers would be dismissed unfairly and 5000 underperforming teachers would be retained. If the administrators used data accumulated over 10 years, "only" about 2,500 teachers would be wrongly categorised. Who would choose to enter a profession in which chances of unfair dismissal were so high? (Interestingly, most of the teachers so dismissed would be teachers in the early years of their career who have not yet acquired the experience necessary for superior results. It would be a strange profession that dismisses its neophytes rather than helping them to become excellent teachers, as many of them would.)

Commenting on this, Wu says, "Essentially the accuracy of labelling teachers as effective or ineffective based on students' test gain scores is not much better than tossing a coin" (Wu, undated, p. 4). This seems an overstatement (the odds are better than that) but she expresses the general thrust of these studies, namely that such scores are seriously unreliable as evidence of teacher quality. It is not surprising, therefore, that after reviewing six major models of VAM, Sloane and colleagues (2013) concluded that "VAM –based estimates should never serve as a single indicator of teacher effectiveness, and high-stakes decisions should never be made primarily on the basis of VAM-based estimates" (p. 64).

2. Just as teachers can be held responsible for only a part of a child's achievement, they can be held responsible for only a part of a child's *rate of improvement*. It should be obvious that outside influences not only affect the child as he or she enters the classroom but also play a major part in a child's *progress*: during one teaching year. Let us say that child A reads the daily paper, consults encyclopaedias, downloads information from the internet and spends holidays in challenging places while child B sits around the house or works in a takeaway bar, uses the computer only for games, and spends the holidays "hanging out" with his mates in the city. How could anyone think it fair or reasonable to ascribe praise or blame to the teachers of these two students for their progress in learning? In the case of secondary school students, the situation is even more dire: can one, for example, ascribe success to the English teacher for progress in written expression when, in fact, it is the student's History teacher who has insisted on good written expression, or praise the Mathematics teacher for progress due to the science teacher's insistence on mathematical precision? Finally, classrooms do not consist of randomly sampled students as would be required for proper monitoring. Some classes consist largely of children who have the motivation and the prior learning which will enable them to make rapid progress. Other classrooms consist largely of children who are lacking in motivation and/or prior learning and who are therefore are capable of making only modest progress. And others occupy every position in between these two extremes.

3. There are problems in selecting the outcomes to be assessed. The ones that are the most easily measured are not necessarily the best for assessing the overall quality of learning (Think of the variety of measures used in the highly regarded NEMP studies in New Zealand which tried to get at deep understanding of science, artistic production, problem solving, etc.). It is, of course, easy to say that the assessors should concentrate on progress in literacy and numeracy, which are to be measured by objective tests. But that is to encourage teachers to ignore other important areas of the curriculum (e.g. science, technology, the arts, and physical education) and to focus on coaching students to pass the tests.

4. There are many other explanations for improved assessment scores: students become more familiar with the assessments; students are coached in taking the assessments; teachers concentrate on the assessed areas and neglect other parts of the curriculum;

parents pay for extra tuition (Bray, 2009); there has been conscious or unconscious manipulation of the results. In the opinion of some, this is already occurring in secondary schools in relation to NCEA results (See Woulfe, 2013).

Despite these strong reservations many politicians and their advisors still believe that VAM is the way to go. What then can further research tell us? Two major studies illustrate some of the problems in devising, justifying and using VAM for assessing teachers and schools.

1. The Economic Policy Institute (EPI) Study (Baker et al., 2001)

This briefing paper was prepared by a group of distinguished scholars, five of whom have been presidents of the American Educational Research Association**.** They begin by pointing out that some American States are now considering plans that would give as much as 50% of the weight in teacher evaluation and compensation decisions to scores on existing tests of basic skills in math and reading. The authors say, "based on the evidence, we consider this unwise" (p. 2). Their arguments for this conclusion are as follows.

1. In study after study VAM scores have been shown to be very unstable. They cite, for instance, a study of five large urban districts, in which 20% of teachers were placed in the top group in the first year. However, less than one-third of these teachers scored in the top group the following year and a further one-third moved down to the lower 40%. The same dramatic fluctuations were found for teachers in the bottom 20% in the first year of analysis (cited, p. 2). Another study found that teacher effectiveness ratings in one year could predict only between 4% and 16% of the variance in rating in subsequent years (cited, p. 2.) In short, the scores gained by teachers in the first year of assessment could not predict even broadly their scores in the second and subsequent years. As will be discussed more fully later (Section 6, Case Study 2), Gorard has conducted many studies of the VAM scheme, which has been used for some years in England (the CVA) and concluded: "CVA is volatile, unreliable, and based on high levels of measurement error" (Gorard, 2012, p. 44). This means, argues Gorard, that "it is not something that can form an ethical basis for policy or practice decisions. Parents should not be (or have been) encouraged to choose, or inspectors to judge, schools on this basis, until the situation is much clearer" (Gorard, 2012, p. 45).

2. Baker and colleagues (2001) argue that what are called “teacher effects” are very likely a complicated set of factors which includes much else beside the actual teaching ability of the teachers being assessed. A study was designed to test this possibility. It used all the methods of VAM after controlling for other factors, but applied the model backwards to see if credible results were obtained. It found that students’ fifth grade teachers were good predictors of the students’ fourth grade test scores. They conclude: “Inasmuch as a student’s later fifth grade teacher cannot possibly have influenced the student’s fourth grade performance, this curious result can only mean that VAM results are based on factors other than teachers’ actual effectiveness” (cited p. 2).

3. The authors refer to the famous campaign No Child Left Behind (NCLB) that, in the interests of promoting achievement for all children, instituted a massive programme of testing in American schools accompanied by sanctions for schools and teachers not delivering the desired results. The NAEP (National Assessment of Educational Progress) evaluation is undertaken on behalf of the federal government with a statistically representative sample in each state. The authors conclude:

> The NCLB approach of test-based accountability promised to close achievement gaps particularly for minority students**.** Yet, although there has been some improvement in NAEP scores for African-Americans since NCLB, the rate of improvement was not much better in the post NCLB period than in the pre-NCLB period, and in half the available cases, it was worse (p. 5).

4. The authors also provide a Table (p. 6) which shows that the progress of white students was lower after NCLB than before in both of the subjects (mathematics and reading) and both of the levels (fourth and eighth grade), while the progress of African-American students improved slightly for two of the tests (fourth grade reading and eighth grade math) and declined in the other two tests (fourth grade math and eighth grade reading). Statisticians might argue about statistical significance but an ordinary reader would be forced to conclude that *this massive and expensive evaluation scheme used in all states and over almost ten years made no real difference to children’s learning*.

5. The authors go on to point out that the reasons such attempts to measure teacher effectiveness fail so dismally is that there are many factors that influence test scores.

Among these are: (i) other teachers of the child (e.g. teachers in previous years or, in secondary schools, teachers of other subjects); (ii) the quality of school resources: class size, curriculum materials, specialist support, library and media resources etc; (iii) the variety of out of school learning experiences at home and in the community; not least of these is the fact that educated and wealthy parents provide extra learning opportunities every week day, at weekends and in the holidays; (iv) Student learning is influenced by the "learning loss" that occurs over the long vacation. Although this occurs for most students, lower-income students lose a lot more than their more advantaged colleagues (p. 3).

6. For all these reasons teachers tend to get lower scores when they teach new English learners, special needs students, and low income students than when they teach the more affluent and educationally advantaged. Sophisticated statistical techniques to "control" for social advantage have largely failed to deliver the goods. (See Appendix B.)

7. It is not surprising therefore that the research community has cautioned that even when sophisticated VAM methods are used, they should never be "used for high stakes decisions such as pay, evaluation or tenure" (p. 2). The authors provide many references in support of such cautions. To select only two: The Board on Testing and Assessment of the National Research Council of the National Academy of Sciences stated: "…VAM estimates of teacher effectiveness should not be used to make operational decisions because such estimates are far too unstable to be considered fair or reliable" (p. 8). And the highly regarded private research firm RAND stated: "The research base is currently insufficient to support the use of VAM for high-stakes decisions about individual teachers or schools" (p. 3).

8. In addition, the authors argue, there are a number of negative consequences that flow from reliance on VAM. The main ones are: (i) It leads to narrowing of the curriculum. If teachers are rewarded for their students test scores in mathematics and reading they will concentrate on these subjects and on the parts which are easiest to drill and to coach. Where VAM flourishes, the arts, language, and other subjects are neglected and children get a weaker form of education. In addition to that, the tests used are very narrow and do not assess important qualities such as problem solving, creativity and communication. It is instructive, say the authors, that between 2000 and 2006, when testing was in full

flight (and some say yielding improved results), the USA results on international tests (which demand more than simple recall) dropped substantially and more and more universities had to set up remedial courses to enable students to cope with higher level study; (ii) Teachers are discouraged from working with special needs students and children from low income homes, and schools (anxious for their reputation) resort to all kinds of strategies to exclude such children; (iii) Teachers are discouraged from co-operating with other teachers on the grounds that they might thereby lose some tactical advantage. At a time when team teaching and sharing professional knowledge are regarded as educationally very important, VAM works against the cooperative spirit.

2. The University of Chicago Study (Borman & Kimball, 2005)

It is frequently claimed that a major purpose of schemes to improve the effectiveness of teachers is to ensure that "tails of underachievement" are directly addressed and the well-known "gap" between the successful and the unsuccessful is eliminated or at least substantially reduced. Borman and Kimball (2005) set out to test the hypothesis that improving teacher quality pays off in relation to closing the achievement gaps. Before reporting the results of their own study, they reviewed the relevant literature. They point out that:

- Research in some local communities has shown that low-income and minority students tend to be taught by underqualified and under-skilled teachers. In contrast, analyses of national data have shown relatively equal distribution of qualified and skilled teachers.
- "There is surprisingly little evidence that highly qualified teachers successfully close the achievement gaps within classrooms that they teach" (pp. 2 & 3). However, Rowan and colleagues (2002) found differences within schools as to the success or otherwise of teachers in teaching disadvantaged students. Although they were unable to identify the characteristics and practices of "good" teachers "the results did suggest that some teachers may be more effective than others in closing achievement gaps" (cited p. 3)

For their own study Borman and Kimball (2005) combined standards-based ratings for nearly 400 teachers and achievement results for over 7,000 students from grades 4-6 in

Washoe County, a mid-sized school district in Nevada. This district uses a scheme based on Danielson (1996) and hence is similar to that in the Cincinnati TES (See Section 6, Case Study 1). They set out to answer two questions:

1. *Is teacher quality distributed equally across classrooms of varying compositions?* Their answer is: "Students from poor, minority, and low-achieving backgrounds have access to teachers of lower quality, as reflected in the teacher evaluations' scores" (p. 17). They suggest two possible explanations for this. The first is that the better teachers may be assigned, and seek to be assigned, to classrooms with more advantaged students. This could help to explain how the system constricts the educational opportunities of children from less advantaged backgrounds. The alternative explanation is also very plausible: that the teachers in the different classroom contexts may actually be of similar quality but "those teaching in less advantaged classrooms may be *perceived* [emphasis ours] by educators as less effective due to the attributes of the students that they are teaching" (p. 18). This explanation ties in with findings from other studies which suggest that the behaviour of students in disadvantaged schools may cause evaluators to rate the teachers lower than is warranted while those of students in more advantaged schools may receive higher ratings based on student behaviour alone. These findings need to be carefully considered before we too readily "blame" teachers for low achievement or "praise" teachers for high achievement. Nevertheless, say the authors, overall, better teachers are associated with slightly better learning outcomes, there being an effect size of between 0.1 and 0.2 (see Appendix).

2. *Is teacher quality associated with improved equality in terms of student achievement?* (Do good teachers close the gaps?) The results of the study are mixed. There are some individual results (e.g. for some subjects at some levels) which suggest positive effects but "in no instance is teacher quality related to closing the gap between low and high achievers" (p. 19) The authors conclude that focused training for work in urban schools may do more to reduce gaps than concentrating on "teacher quality" in some undifferentiated sense.

3. Econometric Studies

There is a further group of studies that arise from the econometric tradition (e.g. Chetty, Friedman, & Rockoff, 2012; Hanushek, 2011; Leigh, 2010; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). Typically, these involve large data sets with in some cases more than a million data points of student achievement. Researchers use these to claim that teachers are significant in explaining differences in performance. Some studies suggest that teacher-specific gains are often sustained when good teachers move schools (Chetty, et al., 2012). It follows, they argue, that (1) some teachers are much more effective than others in raising student test scores and (2) if all students were to have such teachers they would also achieve at the higher levels. (The logical connection between these two claims is dubious.) As a further complication some studies (eg. Chetty, et al., 2012) also claim to show that higher test scores are correlated with higher incomes in later life. At times, indeed, claims verge on fantasy as, for example, "these results imply that the black-white test score gap in Australia could be closed in five years by giving all indigenous pupils teachers at the 75th percentile" (Leigh, 2010, p. 484). Despite serious reservations about the use of econometric methods to provide VAMs for teacher incentive schemes (see Gorard, 2013) these studies are frequently quoted by Treasury officials, political commentators and politicians to justify VAM and Performance Pay. However, this literature itself (e.g. Hanushek & Rivkin, 2010; Rivkin, Hanushek, & Kain, 2005) includes strong reservations about such conclusions. Even where policies are suggested, they are qualified, "While these calculations show that good teachers have great value, they do not by themselves have implications for optimal teacher salaries or merit pay policies" (Chetty, et al., 2012, p. 51). (See Birks, et al, 2013 for an elaboration of these points including a full critique of Leigh, 2010).

While we do not question the central importance of good teachers, it is our view that these kinds of studies, often taken out of context by politicians, bureaucrats and newspaper editors, are of limited help in understanding teacher effectiveness and even less help in improving it. There are two major types of problems with such research: (1) methodological problems; and (2) policy problems.

Methodological problems

- There is the question of persistence. Leigh's claim above is based on simply adding up the effects of five individual years' gains as if all gains persist and additional gains are cumulative. However, we cannot be sure that gains are sustained, with many

studies suggesting that there is "depreciation of learning in educational production functions" (Hanushek, 2011, p. 472). Jacob et al (2008) found that "our estimates suggest that only about one-fifth of the test score gain from a high value-added teacher remains after a single year…. After two years, about one-eighth of the original gain persists. The observed fadeout is comparable for both math and reading, and is robust to several specification checks" (p.33). In this case, it should be noted, a teacher coming in after a student has had a "good" teacher will also have to be "good" just to maintain the student's performance but (if the fade effect is ignored) could be categorised as a merely average teacher. Consequently we cannot be sure that a student who makes large gains in one year will make similarly large gains in subsequent years if assigned to a "high scoring" teacher. There is also an assumption that teachers who are in the top quintile by VAM in one year are necessarily those who are in the top quintile in a subsequent year. Earlier in this Section we have presented evidence of the wide variability in VAM scores for the same teacher over more than one year.

- The econometric models assume that students are randomly allocated to teachers but in the real world this is not the case. Such studies are unable to take into account the non-random allocations of students across schools and teachers across classrooms and the effects of this on the VAM scores of individual teachers. Thus, results may be due to the "mix" of students in particular schools and the distributions of teachers across classrooms within schools. Very often, the data can be explained in alternative ways and may not reveal anything about different teachers at all. Even if the magnitudes are such that successful change seems obvious, these relate to changes in *average scores*. Results from averages may not translate into a meaningful assessment system for individual teachers or particular groups of students. (more about educational policy in (2) below)

- The approach is simplistic. Test scores may not be good indicators of value added because they are a narrow measure and could be the result of many factors besides teacher quality in isolation. Teacher effects are commonly assumed to take the form of a fixed increase in test score for all students. Relationships are commonly estimated on existing data over a limited number of years. Consequently they say little about the ability of teachers to replicate their successes (or failures) at other

times or with other students. Even if test score gains measure the right dimensions of educational attainment, on their own they are useful only if there were to be full persistence of the effects of previous teachers and achievable gains are independent of starting level. Neither of these seems correct.

- The studies rely on the assumption that teachers can be ranked on a single scale of "effectiveness", often implied to be a "normal distribution". Given that all teachers are expected to demonstrate at least minimal skills (i.e. teaching effectiveness), it is safe to assume that teaching competence is *not* normally distributed but rather that it is extremely skewed - in the same way that the effectiveness of medical practitioners is extremely skewed**.** Most practitioners (medical and educational) are at least adequate and many are far better than that. This error seems to arise from the fact that the studies typically use standardised tests of reading and mathematics which have been specifically designed to produce a distribution of student scores resembling a normal distribution. Thus the "normal distribution" of student achievement scores (an artefact of standardised tests) is assumed to mirror a "normal distribution" of teacher effectiveness which has not been shown and is highly implausible.

- The fact that the assumptions underlying the econometric studies are not widely understood can result in their conclusions being misused. For example, the association with improved incomes cannot be directly attributed to teachers (or school) at all. They may be explained by the correlations between IQ and income, family background and income and college attendance and income. Statistical controls may not be sufficient to guard against these alternative explanations (see Appendix). Of course, such methodological issues are highly controversial and need to be threshed out by experts in econometrics. They cannot be dealt with adequately by educational researchers and certainly not by politicians who, to their shame. often put out short "authoritative" sentences with no references that can be checked eg the Minister of Education, Hon Hekia Parata, solemnly told a conference of private schools that, "experts had found that four consecutive years of quality teaching eliminated any trace of socio-economic disadvantage" (Fea, 2013). This statistical nonsense was not criticised by the media—to their shame too. Jacob, Lefgren, &

Sims make clear that it is not possible to add up consecutive gains, so we should reject claims that "a series of high value-added effects for a hypothetical student with a string of good teachers may be simply added together" (2008, p. 33). (See Birks, et al., 2013, for a fuller discussion of such claims.)

In addition to the problems of the methodology, there are other problems in using this kind of research for making educational policy.

Educational policy problems

- While Hanushek and others in this tradition insist that teachers are the most important school influences on learning, they also concede that "it has not been possible to identify any specific characteristics of teachers that are reliably related to student outcomes" (Hanushek, 2011, p. 467). That is to say, these researchers *can tell us nothing about what constitutes a good teacher*. They provide no guidance as to how to recognise, select, or train a good teacher. For the purposes of educational policy, the research is useless.

- Proponents indeed seem to recognise this and hence the "fallback position" is that all that we can do is to eliminate (i.e. "dismiss") the least effective teachers. As we showed earlier, getting rid of the lowest achieving teachers would lead to large numbers being dismissed unfairly since VAM data are unreliable. Undoubtedly, such a crude policy "setting" would produce serious adverse consequences for teacher recruitment, morale and retention; and for students' classroom learning environments and relations. These "interaction effects" are ignored or obscured by the simplified relationships assumed in the econometric models.

- Although large studies may give *statistically significant* results on teacher fixed effects overall, this may be due to the large number of observations. The effects may not be large enough to be worthwhile from policy making. Furthermore, the results may not translate into a meaningful assessment system for individual teachers or particular groups of students. (Birks et al, 2013)

- It may even be the case that “teaching effectiveness” (as “measured” in these studies) has little to do with teachers at all in the sense of individual performance. What may be showing up in the studies are some interactive effects involving student ability, teacher performance, peer effects, and much else besides. This indeed is what Baker et al. (2010) argue. According to them “teacher effectiveness” is not a “stable construct” that can be measured and they suggest that what are called “teacher effects” are very probably a complicated set of factors which includes much else beside the actual teaching ability of the teachers being assessed (p.13). (See our discussion in (1) of this section).

- The measure of teacher quality used in these studies is a limited one: achievement on objective tests of reading and mathematics. However, teachers in primary schools are expected to cover the eight areas of the curriculum and they are all important. By using only tests of literacy and numeracy, the researchers ignore those teachers who “turn students on” to a love of literature, music or art and who produce students who are critical and creative. It is ironic that while business leaders ask schools to produce flexible, creative and cooperative employees, so much of current research and policy is based on performance on narrow tests of basic skills which emphasise the very opposite traits.

- A further criticism of the narrowness of scope in these test-result approaches to teacher assessment is made by Gorard, who concludes that, “with our current approaches based on pupil test outcomes we cannot safely identify any individual teacher who is differentially effective with equivalent pupils” (Gorard, 2013, p. 80)

- Those who conduct such studies and propose policy changes rarely calculate the cost-benefit ratios of the policies that they favour. It may be, for example, that if money required to establish a VAM regime was applied, instead, to professional development in more effective teaching techniques that this would have a much great effect on student achievement.

- New and untried policies would alter the environment and may therefore change the underlying structure from that which the policies have been derived. This

phenomenon has been described more generally (Lucas, 1976) and is now referred to as the "Lucas critique". Chetty et al. (2012, p. 1) acknowledge this problem: "Using VA in teacher evaluation could induce counterproductive responses that make VA a poorer measure of teacher quality, such as teaching to the test and cheating".

- Even if these studies show that some teachers are better than others (which common sense would suggest is right), they are of little use to policy making unless they can show *which teachers are in fact better, and why*. Only then can we begin to develop the education programmes that will produce more of these highly effective teachers. Sadly, however, as we have shown in Section 3 on teacher effectiveness, this is something we do not yet know how to do. This is a particular problem when some of the weaker teachers are actually the new ones who are building up experience and will normally improve. It would be very strange to "sack" teachers who are acquiring the very experience which this research itself shows is a significant criterion of the "good teacher." Reducing this to its logical absurdity, we would have to say that no teacher should be employed unless she/he has had more than 10 years' experience!
- Such conclusions seem obvious since all such policies are negated by the fact that in econometric studies "best teachers" cannot be identified, except on subsequent student test scores which rests on a tautology: "good teachers are those teachers who produce good results".

SUMMARY

This section has shown that:

- In recognition that teachers cannot be held accountable for effects outside their control (e.g. home background, previous knowledge, classroom composition, natural ability and mental engagement), researchers have turned to trying to assess what "value" a teacher has "added" to the student's learning over the time they were in her classroom. This is a particular form of achievement based assessment known as value added measurement (VAM). VAM is problematic because: (1) it takes little account of differential student abilities; (2) it is difficult to control for the widely different learning experienced by students outside school; (3) the variables which are the most easily measured are not necessarily the best for assessing the overall quality of learning; and (4) there are many other explanations for improved assessment scores such as the effects of other teachers and the quality of school resources.
- VAM scores have been shown to be very unstable, with dramatic fluctuations from year to year, and high rates of measurement error. Consequently, the scores gained by teachers in the first year of assessment cannot predict even broadly their scores in the second and subsequent years.
- There are also negative consequences that flow from reliance on VAM. The main ones are: (1) it leads to narrowing of the curriculum; (2) teachers are discouraged from working with special needs students and children from low income homes, and schools (anxious for their reputation) resort to all kinds of strategies to exclude such children; and (3) teachers are discouraged from co-operating with other teachers on the grounds that they might thereby lose some tactical advantage. At a time when team teaching and sharing professional knowledge are regarded as educationally very important, VAM works against the cooperative spirit.
- Econometric studies suggest important connections between teacher qualities and student achievement on tests, but they face serious statistical and analytical problems and are unable to provide any analysis of what constitutes a good teacher or any advice to policy makers regarding teacher recruitment or training.

SECTION 6. HIGH STAKES ASSESSMENT OF TEACHERS

> *VAM–based estimates should never serve as a single indicator of teacher effectiveness, and high-stakes decisions should never be made primarily on the basis of VAM-based estimates (Sloane et al, 2013, p. 64).*

Despite all of the problems discussed in Section 5, the creation of new and "high stakes" assessments of teachers seems to be common in recent proposals for school reform and we have shown that the New Zealand Treasury (and, perhaps, the Government) is keen to promote such assessment here. Thus, in this section we look at two such assessment systems to see if there are lessons for New Zealand: we look at the Teacher Evaluation System (TES) in Cincinnati, United States of America, and the Contextualised Value Added (CVA) scheme in the United Kingdom.

The notion of "performance" here is quite ambiguous. On one model, the teacher is evaluated on what he or she does (the performance, like an actor). On the other, the teacher is evaluated in terms of "how successful he or she is" (the performance in raising the "achievement" of his or her students). This dichotomy is marked by different approaches to the measurement of "performance". In the first model (instanced by the Cincinnati TES), the teacher is rated on how she "performs" (acts), without regard for student learning. In the second model (instanced by England's CVA) attention is focused on the extent to which students learn. As we shall see, both models are problematic.

1. Case Study One: United States of America

The Cincinnati Public Schools' Teacher Evaluation System (TES)[7] is recognized as a significant model of teacher assessment and one which is essentially replicated in many American states. All teachers participate in a Comprehensive Evaluation at defined intervals - the first year as a new teacher, the fourth year, then every five years after that point. The Comprehensive Evaluation consists of an orientation meeting to learn about the evaluation process, a readiness conference with the evaluator to share details about the teaching assignment, followed by at least four classroom observations. An Annual Assessment, one

[7] This TES overview is largely a summary of the information provided on the Cincinnati Public Schools website at http://www.cps-k12.org/about-cps/tes

classroom observation conducted by the teacher's administrator, occurs during those years a teacher is not scheduled for a Comprehensive Evaluation.

This scheme was introduced in 2001. Evaluators are high performing, experienced teachers who have taught in different schools in the district. They are supported by groups of administrators. Teachers are observed four times in the year: three times by the peer evaluator and once by an administrator. Both the evaluators and the administrators complete an intensive TES training course during which they are expected to accurately assess videotaped lessons.

The TES is based on the framework prescribed by Charlotte Danielson in her book *Enhancing Professional Practice: A Framework for Teaching* (Danielson, 1996). This framework divides skills and responsibilities into four domains: Planning and Preparing for Student Learning, Creating an Environment for Student Learning, Teaching for Student Learning, and Professionalism. In each of these domains good teaching is assessed against sixteen standards which articulate the skills and practices believed to be integral to good teaching**.** These standards establish clear expectations for performance and professional development in the Cincinnati Public Schools. Overall, there are four levels of expertise (see below).

Teaching performance for each standard is described in a scoring guide. These standards and rubrics also have been adapted for specialists in the school (librarians, counsellors, social workers, school psychologists, etc.).

The scheme was developed from consultations between the district authority and the teachers' unions. It was recognized that evaluating teachers was not enough: there must additionally be procedures to improve teacher quality and help teachers experiencing difficulty. As a result, the Peer Assistance and Evaluation Program (PAEP) was developed to help new teachers and experienced teachers having difficulties.

In addition, it was recognized that many high-quality teachers were leaving the profession. As a result, the Career-In-Teaching Program was developed to provide incentives to attract and retain quality teachers in the profession, to improve and encourage teachers' professional growth opportunities, and to support teachers to improve student achievement.

The Career-In-Teaching program identifies five teaching levels, based on the teacher's evaluation scores. Teachers work toward achieving higher status which involves more responsibilities and extra pay.

- Level One, Apprentice: A teacher without previous teaching experience. This level prepares teachers to pursue a career in teaching.
- Level Two, Novice: A teacher who has met licensure requirements and is working to develop the skills required for a career in teaching.
- Level Three, Career: A teacher who has demonstrated the skills needed to have a career in teaching.
- Level Four, Advanced: A teacher who is continuing to master the art of teaching, demonstrating a distinguished level of teaching.
- Level Five, Accomplished: A teacher who has demonstrated outstanding teaching.

The Peer Assistance and Evaluation Program has two major roles.

1) Through its Apprentice Component, it assists teachers in their first year in the Cincinnati Public Schools by helping them refine their teaching skills and orienting them to the district, including its goals, curriculum and structure. Through this component, each teacher is evaluated and assisted by a consulting teacher.
2) Through the Intervention Component, the program assists experienced teachers who are judged to have serious instructional deficiencies. When a teacher has not met the expected performance standards, the teacher can be referred for intervention. A joint panel of teachers and administrators reviews the referrals and assigns consulting teachers to work with those teachers to improve their instructional skills and to improve the teachers' levels of performance. In cases where improvement does not occur, the panel may recommend a second year of intervention or the non-renewal of a teacher's contract.

Lead teachers: A lead teacher helps another teacher improve her skills. The primary role of a lead teacher is to support quality instruction. In addition to having been classified as an advanced or accomplished teacher, a lead teacher must demonstrate leadership in his/her profession, effective communication skills, a consistent pattern of professional growth, cooperation and collaboration, and commitment to teaching as a career. Teachers may apply to be a lead teacher by going through a three-phase application process. A joint board/union

panel grants lead teacher status. Once a teacher is certified as a lead teacher, he/she may apply for lead teacher positions throughout the district. Lead teachers serve at both the school and district level in various roles (consulting teachers, teacher evaluators, curriculum specialists, subject leaders, team leaders and program facilitators, etc)

It is important to note that the Cincinatti TES does not involve any measurement of student learning and hence is in *no way a "value added" model*. It is a "standards" based procedure which is intended to examine and refine the activities of teachers. However, since the point of teaching is to help students learn, it seems important to investigate whether the expensive and time consuming scheme actually leads to improved achievement in students. With this in mind, we now look at two evaluations of the Cincinnati TES.

Taylor and Tyler (2011)

Taylor and Tyler (2011) evaluated the Cincinnati scheme and compared teachers before and after the assessment reforms. They examined data which might indicate whether this scheme of teacher evaluation led to better achievement in learning for their students. Indeed, they claim that their study was the first to look at the claim that teacher evaluations could have a lasting effect on student learning. For a number of reasons they focused on teachers who were teaching 4^{th} to 8^{th} Grade in the years 2003-4 through to 2009-10. They correlated teacher scores in the Cincinnati TES with student achievement in mathematics and found that:

- In line with many other studies, more than 90% of teachers scored in the top two categories. (The authors call this the "leniency effect" but of course the figure instead could be used to claim that "most teacher are very effective" because value judgments are significant in interpreting data.)
- Teachers were more effective in raising student mathematics scores during the year in which they were assessed than they were previously and were even more effective in the years following evaluation.
- A student instructed by a teacher who has been evaluated will score about 0.11 of a standard deviation higher in *mathematics* than a similar student taught by the same teacher before evaluation. They instance two students who began studies with the same teacher in different years both of whom were at the 50^{th} percentile: the student

who went through after the TES would score 4.5 percentile points more than the student who went through before the teacher faced the TES.

- The effects were largest for teachers who initially were judged lower and who had experienced most difficulty in raising achievement scores.
- There was no evidence of any similar impact on the *reading scores* of students.

Several problems are apparent in this study. First, the design of Taylor and Tyler's study compares the results of a cohort of students from one year with a cohort of students from previous years for a given teacher. Such a design cannot control for variation in ability between the cohorts. Within the study the authors claim that the average effect size on student achievement subsequent to their teacher going through the Cincinnati evaluation is approximately 0.11 of a standard deviation in mathematics (or 4.5 percentile points for a median student). From this we can estimate the standard deviation of students' achievement in mathematics to be approximately 40 percentile points. This relatively high standard deviation indicates a large variation in student ability in the background population. As such, we might reasonably expect that there will naturally be noticeable differences in average achievement from year to year, irrespective of the quality of teaching. Moreover, it is common for test scores to increase as the teacher becomes more familiar with the requirements of a given test and better at preparing students and targeting teaching to these requirements. As such, it is unclear what has caused the relatively small effect size of 0.11. It may simply reflect a combination of the natural variation in ability among students in a given class and more "targeted" teaching rather than "better" teaching per se. *It is worth noting that in contrast to the effect on mathematics achievement, the authors found no impact on reading performance.*

Second, the authors are surprisingly lacking in self-evaluation of their findings: the positive scores in mathematics are lauded while the nil results in reading are explained away. Once again it is essential to point out that correlation is not causation. The very modest improvement in student test scores (in maths) may be due to the same cause as the development of the performance measures: political pressure, for example. It is possible that the performance rating had no effect at all. (See Appendix**)**

Third, is an effect size of 0.11 a large effect or a small effect? This is not a matter of mere statistics but an important argument about whether an average gain of some 4 to 5 percentage points is educationally significant. Put bluntly, is this gain in maths scores really worth all the time, resources, and effort (including a pervasive distrust of teachers)? Might it be better for example to give more attention to improving the methods of teaching reading or to extending activities in which mathematics is embedded? Is it justifiable to attach so much weight to an expensive and time consuming method of improving teaching? Thus, as so often, the assessment of teachers rests not just on research findings but on value judgements as to what overall is most worthwhile.

Finally, no evidence is sought about possible "side effects" of all this concentration on the mechanics of teaching. Might teachers lose some spontaneity and something of the individual flair which adults often associate with teachers who influenced them?

Milanowski (2004)

Milankowski (2004**)** reports on an analysis which he carried out of the relationship between teacher evaluation scores and student achievement scores on tests of reading, mathematics and science. Using a "value added framework" he correlated the differences between predicted and actual students achievement in each of these subjects in Grades 3 to 8 with teacher evaluation ratings. As a result he reported that:

1. "Small to moderate positive correlations were found for most grades in each subject tested".
2. When these correlations were combined across grades within subjects the average correlations were 0.27 for science, 0.32 for reading, and 0.43 for mathematics.
3. Thus, he claims, his "results show that scores from a rigorous teacher evaluation system can be used….as the basis for a performance-based pay system" (p. 1).

As with Taylor and Tyler, our reading of the study report identifies several problems. First, we are struck by the lack of critical appraisal of his own findings and the speed with which the reported positive findings (small as they are) are immediately used to support performance pay and "other resource management purposes" (p. 8.) while, at the same time, the author admits that the "results presented here should be regarded as suggestive rather than definitive since they are based on only 1 year of student achievement indicators" (p. 8).

Second, the procedure by which he combines results for each subject over five grades is cause for concern when one considers what science or mathematics looks like in Grade 3 compared to Grade 8 and when we note that only "small to moderate results" were found *within* grades. He acknowledges without comment that "there was substantial test score variance at the teacher level (ranging from 6% for Grade 8 science to 28% for Grade 4 science.)" (p. 6). Is aggregating widely variant scores a justifiable move? And can we be sure that "science" is the same entity in the different grades?

Third, where the district uses all four domains of teacher competence to award pay, this researcher aggregated scores from each of the domains to produce an overall score. This was then used "as an overall indicator of teacher performance" (p.8). Is such aggregation justified or might it distort the actual results?

Fourth, decisions seem to have been made to exclude some teachers and some students from the study. Having begun with a teaching pool of more than 700 teachers he finally reported results for 212. It should also be noted that "a substantial proportion" of students enrolled in each grade in 2001-2002 could not be included in the analyses because grades were missing for one or both years and because of substantial mobility between schools.

Finally, the author claims that "the teacher assessment system was able to identify which teachers had students with higher than expected levels of achievement, as measured by test scores, to a degree greater than chance" (p. 49). However he goes on to assert that "this supports the use of the assessment system as a basis for teacher evaluations …". This conclusion simply cannot be justified. The correlation between the evaluation system and test scores was only 0.3 or 0.4, indicating that only 9 % to 16 % of the variance of student achievement across classes was due to variation in assessed teacher performance. In other words, 84 percent or more of student achievement was unrelated to the teacher assessment measures. When performance pay is based on such an unreliable measure, the potential for serious injustice is obvious**.**

On the basis of the studies reported we are not convinced that the studies reviewed support the claim that the Cincinnati TES predicts student achievement; hence its use as a basis of performance pay is unethical. Of course, it can still be argued that serious attempts to

improve professional *practice* are desirable, provided they do not damage the autonomy and cooperative ethos necessary for good teaching.

2. Case Study Two: England

The Contextualised Value Added Scheme (CVA) is a scheme by which the English Department for Education and Skills assesses the progress of pupils and the achievement of schools. It is intended to show the progress children have made while attending a particular school. Unlike statistics that merely report actual assessment performance, contextualised value added attempts to take into account those circumstances of children's lives that are beyond the school's control. This is done by comparing a child's improvement against that of children with similar prior performance levels and similar circumstances. There are three levels; Level 1 CVA measures the improvement of primary school age children and schools. It measures the improvement of pupils between the end of Key Stage 1 and the end of Key Stage 2. It is based around a median score of 100. Level 2 CVA measures improvement of secondary school age children and schools (i.e. between the end of Key Stage 2 and the end of Key Stage 4). It is based around a median score of 1000. Level 3 CVA measures performance of post -16 pupils from the end of Key Stage 3 to the end of Key Stage 5.[8] Like Level 2 CVA, it is measured around a median score of 1000. CVA takes into account nine factors that are known to affect the performance of children, but are outside of the control of the school. The factors are:

- Gender
- Special Education needs.
- Eligibility for free school meals
- First language
- Whether pupils move between schools
- Ethnicity
- The age (i.e. the month they were born in) of different pupils within the year group
- Whether a pupil has been taken into care (e.g. foster care) at any stage
- The level of deprivations in the area the pupil lives (using the Income Deprivation Affecting Children Index).

[8] Key Stage 1 ends at the age of 7; Key Stage 2, at the age of 11; Key Stage 3, 14; Key stage 4, 16; and Key Stage 5, 18.

It is clear that this measure tries very hard to control for confounding factors which make value added schemes so unfair (see Section 5) beyond the initial achievement level which is usually controlled in VAM models. (See Appendix B) However, there are concerns associated with the statistical method. For example, care needs to be taken when making comparisons involving small score differences. In particular, CVA has a margin of error and the degree of uncertainty in the score increases as the size of the cohort decreases: a different set of students at the same school will almost certainly produce differing CVA scores. The fact that Contextual Value Added scores are averaged across the country also means that it is not possible to isolate local variance.

Two very different criticisms have been made of the scheme.

Firstly, there have been *political criticisms*. Conservatives protested that the scheme is not fair since it "discriminates" against "elite" schools (CVA data suggest that so-called elite schools generally do not do as they claim to do, but simply capitalise on advantaged students). They also argue that the scheme provides "excuses" for schools which fail to improve the attainment of students from low income homes and thereby entrenches low aspirations for such children. Liberals/progressives, on the other hand, have argued that CVA *underestimates* the problems arising from deprivation since many poor people are unwilling to apply for special treatment (free lunches and the like). As we shall see below, it is likely that the government has been influenced by at least some of these political criticisms.

Secondly, there have been *criticisms from researchers* who have examined patterns of examination results to determine whether the evaluation scheme is reliable and valid. Gorard (2006, 2008) examined CVA data in various areas of the country and concluded that CVA is only marginally more trustworthy than raw value added scores and may in fact tell us *nothing* about good teachers or high performing schools. In reply critics have argued that since many schools have *consistent* scores they must mean something. So Gorard set out to examine this proposition by looking at the secondary schools of Worcestershire. From 2006 to 2010, 30 secondary schools were listed as being in Worcestershire. Given that the CVA results for Worcestershire are slightly above the national average (1000) he expected a reasonable number of these schools to score above 1000 and so it turned out. The question then was: "How far away from 1000 does a score have to be before it is significant?" On the basis of a statistical model he estimated that if the scores were perfectly random he would

predict that two schools of the 30 would have a consistent positive score and two would have consistently negative scores. If on the other hand, CVA were a valid measure of a relatively stable characteristic of school quality, he would predict that more schools would have consistent scores, positive or negative. As it turned out, two schools were consistently positive and two consistently negative: this, he argues, is the same result as would be expected of 30 schools if the results are purely random (Gorard, 2012). He argues that there is no reason to expect any different sort of result if the school results were examined "in any other area" (p. 44). He concludes "CVA is volatile, unreliable, and based on high levels of measurement error" (p. 44). This means, argues Gorard (2013), that it is not reasonable to use CVA results for practical purposes: parents cannot rely on them when choosing schools; school leaders cannot rely on them to judge the effectiveness of teachers or departments; and officials cannot rely on them to make decisions about the quality of education delivered in schools.

The Conservative Party, when in opposition, began to express disquiet with the CVA scheme and soon after the 2010 election, the Coalition government announced that the system would be abolished on the grounds that "parents do not understand it". Performance tables will continue to show how much progress pupils make compared to their prior levels of attainment, and how many pupils there are with special educational needs in each school. League tables will continue to be published but they will no longer take account of pupils' deprivation, ethnicity and other background factors when comparing school exam and test results.

Unions have warned that the abolition of the contextual value added (CVA) measure without a replacement will further stack the odds against head teachers and teachers who choose to work in the most deprived areas and may create recruitment problems in schools battling the highest levels of social disadvantage.

The independent think-tank, DEMOS, has recently produced a damning criticism of the whole inspectorial and assessment system in England (Park, 2013). It uses research to argue that, as a result of 20 years of high stakes testing and league tables, the following dire situation has been created:

1. Students increasingly pursue qualifications which are more helpful to the school's reputation than to their lives;

2. Students' intellectual development is forced into a mould which leaves many students feeling failures (schools sometime display colour coded lists of students who are proficient or failing);
3. Schools are unable to develop character skills needed in society. Business is complaining that students lack social skills since these are not tested and hence not encouraged;
4. Teachers are retreating into "safe" approaches (such as lecturing and drilling), fearing that innovative teaching will be marked down;
5. Many children are feeling worthless as a result of judgments made about them.

As a result, politicians are endlessly tinkering with the system in their attempts to improve accountability. On the contrary, the authors of the report conclude:

> The evidence presented in this report so far would suggest instead that the failings of the school system might be the *consequence* of the accountability and assessment system we have. Nothing can be achieved by continuing to tinker with it. If we want to address the problem, only a radical overhaul will do it (p. 52).

They go on to recommend radical changes to the system which are targeted to the English local characteristics but, basically, they involve recognition of the complexity of education and the need to adopt a broader view of the nature of teaching than the one which has been favoured for some twenty years and which, in their view, has failed. This theme will be pursued in the next section as we move towards a more positive view about what we think needs to be done in New Zealand.

SUMMARY

This section has shown that:

- Two specific teacher performance evaluation schemes were examined for the purposes of this report, the Teacher Evaluation System (TES) in Cincinnati, United States of America, and the Contextual Value Added (CVA) scheme in the United Kingdom. The notion of 'performance' in each is quite ambiguous. In the TES, the teacher is evaluated on what she does (her performance, like an actor). In the CVA, the teacher is evaluated in terms of "how successful she is" (her performance in raising the "achievement" of her students). This dichotomy is marked by different approaches to the measurement of "performance".
- The Cincinnati TES does not involve any measurement of student learning and hence is not a "value added" tool. It is a "standards" based procedure which is intended to examine and refine the activities of teachers through multiple structured classroom observations by trained evaluators who are experienced classroom teachers. Comprehensive evaluations are undertaken at approximately five yearly intervals for experienced teachers and more frequently for newer teachers. In addition, a scheme is in place to assist new and experienced teachers who experience difficulties. Two studies have claimed to identify increases in student achievement as a result of the scheme's introduction. However, the reported student achievement gains have been modest, inconsistent across learning areas, and could equally well be explained by other variables not considered by the study authors.
- The English CVA sets out to show the progress children have made while attending a particular school. Unlike statistics that merely report actual exam performance gains, contextual value added attempts to take into account the circumstances of children attending the school that are beyond the school's control. This is done by comparing a child's performance with that of children with a similar prior performance and similar circumstances. Numerous variables are included in the model to attempt to "control" for factors beyond the school or teacher's control. Conservative politicians have claimed that the statistical methodology discriminates against socially advantaged schools, while liberal/progressive politicians have claimed that the CVA underestimates the challenges faced by disadvantaged schools. Researchers have claimed that the CVA suffers from serious reliability and measurement errors to the extent that the results are only marginally more reliable than raw data.

SECTION 7. PERFORMANCE PAY

> *These studies indicated that the PRP [Performance Related Pay] schemes had a limited impact upon employee motivation but that there were a number of negative impacts including deleterious effects on cooperation and team working and divisiveness in the workplace. These were views shared not only by employees but also by their managers* (Farrell & Morris, 2009, p. 80).

Parents and politicians are interested in ensuring that our schools do the best job that they can with the resources available. There are, however, no simple ways of improving a nation's schools. This is because the quality of a school system depends on government policies and funding levels for schools, teacher education provision; and research into teaching and learning. It also depends on the content of the school curriculum, the quality of the teaching materials available, and matters beyond the control of government such as the proportion of families who have the income and parenting skills to take advantage of the schooling which is available to them.

One of the changes designed to improve school effectiveness that is currently being considered by the Government is the use of value added measures of student achievement as the basis for some system performance based promotion for teachers.

1. Theoretical Issues with Performance Pay

The proposal to pay teachers for their "performance" is suggested by research into motivation where it has been observed that paying employees for the results of their efforts can have the effect of motivating improved performance – at least under certain circumstances. For example, sales staffs are often paid a bonus on top of their base salary for exceeding a targeted level of sales over a given period. This section addresses the question of whether performance pay could be used to motivate more effective teaching on the part of classroom teachers. To even begin answer this question we need to recognise that:

- Primary teachers are responsible for student progress in eight essential learning areas. Student achievement must be assessed in all eight areas otherwise the performance

payment would have the effect of limiting teaching to just those achievement areas which are being assessed.

- It is difficult to measure student achievement accurately because student performance is affected by variations in student attitude, motivation, and stress level at the time of testing.
- To be fair, all tests would need to be administered by an agency which was independent of the school in which the teacher was working.
- A performance payment based on a single end of year measure of student achievement in each area would never be fair because end of year achievement is a function of multiple influences including: (1) student ability; (2) home support for learning; (3) level of student achievement at the start of the year; (4) learning arising from the student's own efforts; (5) other influences such as other teachers and coaches; and (6) the work of the individual teacher.
- It is often argued that the problems can be avoided by measuring *improvement.* (Improvement measures are sometimes referred to as "value added" measures) A performance payment which depended on measures of *improvement* in achievement (e.g. from the beginning of the year to the end of the year) would not be fair since the teacher is only one influence on student progress (see Section 5).
- Teachers work as a team helping and supporting each other within a well-organized school so at least some of any performance payment would need to be adjusted to recognise this.
- Any system must also recognise differences in school decile level, student ability and the overall composition of the school.
- Any attempt to introduce performance pay for teachers would need to be preceded by a very carefully controlled cost-benefit analysis. Cost-benefit analyses may well show that there are alternative policies which could be used to lift teacher performance for a smaller investment than would be required to set up and maintain a performance pay system.

2. Research on Performance Pay Schemes

A major review of research on performance pay was undertaken by the Australian Council for Research in Education by Ingvarson, Kleinhenz and Wilkinson (2007). Its purpose was

to: (1) review pay and bargaining arrangements for teachers in Australian schools; (2) identify any policy and legislative frameworks that may impede the introduction of performance pay; (3) provide an overview of "stakeholder attitudes to performance pay"; (4) review recent international research on the impact of performance pay on a range of intended goals such as "higher levels of teacher retention, improved teaching standards, improved student outcomes, recognition, etc."; and (5) identify gaps in research on any of the above (p. 127).

The main conclusions of the review were as follows. First, to be valid and reliable, judgments about teacher performance for "high stakes" purposes such as registration, reward, career advancement and promotion must be based on several sources of evidence. Second, "standards" are now commonly used to specify what teachers should know and be able to do. Such standards are multidimensional (e.g. knowledge, skills, dispositions), and therefore multiple forms as well as multiple sources of evidence are required. Third, valid and reliable assessment of teacher performance additionally requires "multiple, independent trained assessors of that evidence" (p. 6) together with evidence about the context in which judgments about performance are being made.

While we cannot here reprise all the issues canvassed in their report, we are acutely aware of the New Zealand Treasury's advocacy for the development of value added measures of student achievement as part of a revised individual teacher appraisal scheme (Treasury, 2011b). In this regard we note Ingvarson and colleagues' comment that "the consensus among those who are familiar with these [value added] schemes is that they do not provide, and are unlikely to provide, a valid basis for decision-making about the quality of teaching" (2007, p. 60). Recent studies of attempts to introduce "value added" forms of teacher recognition and reward in the United States reveal significant problems of reliability and validity. Some studies report considerable variation in scores for the same teacher from year to year that are attributable to high error rates in testing (Wu, 2012; Gorard, 2013), others demonstrate a weak correlation between value added measures of student achievement in mandated tests and other measures derived from classroom observation of teaching (Buris & Welner, 2011) or from alternative forms of assessment (Rothstein, 2011). Overall, despite claims of their advocates for the reliability and validity of the measures used, serious issues have been raised with regard to the use of student achievement tests to make judgments

about the performance of teachers and/or the value they add to their students' learning (See Sloane, et al, 2013).

Equally, with regard to the Ministry of Education's (2012) advocacy of "recognition and reward" schemes and career pathways for teachers that are based on a combination of standards, competencies and professional learning and development, we note Ingvarson and colleagues' caution that developing a valid basis for evaluating teacher performance "is by no means a straightforward task as it may seem" (p. 63). This is because:

> Professional work is not as amenable to simple measures of account as that of many other occupations. Rather than outcomes, professionals are held accountable for applying practices consistent with current research and best practice rather than one-dimensional outcome measures. (p. 62)

To this we would add our own caution that the quality of research on teaching is known to be highly variable, while models of best practice derived from research are too frequently normative and merely comprise abstract lists of potentially useful pedagogical ingredients, as it were, not a recipe for the application to practice. Current research and best practice would therefore provide an equally problematic basis for assessing teacher performance because the quality of teaching also depends both on the particular classroom context in which it is enacted and the interaction of numerous explicit and implicit variables to do with both teaching and learning.

In terms of their analysis of recent research on performance pay schemes, Ingvarson and colleagues note that such schemes define performance very differently.

> Proponents of some schemes argue that standardised tests of student learning outcomes should be the main indicator of teacher performance. Others argue that performance should focus more directly on evidence about what students are doing in classrooms as a result of conditions for learning established by teachers. (p. 13)

In our view, teachers may reasonably be held responsible for exercising their professional knowledge, skills and dispositions to create supportive conditions for students to learn in their classrooms. However, teachers cannot in any sense control or guarantee what students

learn, or how much – yet this is clearly implied in any scheme which measures student learning outcomes or gains in learning outcomes over the period of time spent with the teacher. As Ingvarson and colleagues comment, a distinguishing feature of professional work is that "practice depends on the application of expertise and judgment to what are frequently non-routine problems" (p. 62). For example, judgments about medical practitioner performance are necessarily about whether or not accepted standards of good practice were followed, not simply patient outcomes. Clearly, "variation in outcomes is not necessarily because of variations in professional performance" (p. 62), yet in our experience this is all too often forgotten when schemes for the assessment of teachers' performance are proposed.

Ingvarson and colleagues identified three main types of performance pay schemes in the research literature: "merit pay; knowledge and skills-based; and certification approaches" (p. 13). Much of the evidence on *merit pay* dates from the 1970s and 1980s in the United States. Broadly, the approach failed to meet the necessary evidentiary validity and reliability criteria specified by Ingvarson and colleagues, which led to staff dissatisfaction. Moreover, teachers were typically evaluated against each other for a limited pool of funds, usually in the form of individual bonus payments. In contrast and somewhat surprisingly given the social nature of much of teachers' work outside the classroom, there is limited empirical research on the effects of group or institutional reward schemes – see for example Lavy (2002). Another challenge to identifying and rewarding good teachers in this way is establishing consistency of teacher performance. For example, one retrospective analysis of student achievement where students had been randomly allocated to classes reported different gains in student achievement for individual teachers in different core subject areas (mathematics and reading), and at different career stages (Dee & Keys, 2004).

Knowledge and skills-based schemes traditionally rewarded teachers for acquiring additional tertiary qualifications or short-course attendance credits. Newer schemes are based on the development of differentiated teaching standards and technologies for assessing these. "These new approaches may still include interviews, classroom observation and student evaluations, but they may also include portfolio entries containing videotape evidence and evidence of improved student knowledge and skills over time" (Ingvarson et al, 2007, p. 14). Clearly, to enjoy credibility with and support from teachers, all those who meet the relevant knowledge or skill acquisition standards must receive the reward.

Certification schemes are characterized by the involvement of a professional body that operates independently from the teacher employing authority. The role of the professional body is to develop professional standards of teaching competence and/or practice, to assess individuals against these and to certify acceptable performance. Professional bodies may be system-wide or specialized (e.g. subject associations).

Looking across evidence from studies of the three types, Ingvarson and colleagues conclude that "performance-based pay systems are more likely to have a positive impact when their development and operation is seen as a mutual responsibility between employing authorities and professional associations" (p. 16). This is because the two groups are argued to have "complementary roles" (p. 16).

In summary, the authors state that performance-pay schemes are more likely to be successful if the following criteria are satisfied.

a) Their guiding purpose is to give substantial and valued recognition to teachers who provide evidence of professional development to high teaching standards (which includes evidence of student learning outcomes);
b) Valid (research-based) standards have been developed by expert teachers in their specialist field of teaching to provide long-term goals for professional development;
c) Appropriate research has been completed to develop reliable and valid procedures for gathering evidence to indicate whether teachers have met those standards;
d) The assessment of performance procedures are conducted by an agency external to the school to ensure reliability, comparability and fairness;
e) Teachers have adequate opportunities to learn the knowledge and skills required to put the standards into practice;
f) A teacher's ability to demonstrate that they have met the relevant standards leads to valued professional recognition, enhanced career opportunities and significant salary increases;
g) Teachers who reach high standards of performance gain access to interesting, challenging and well-supported positions in schools where they can provide leadership to improve teaching and learning; and

h) Governments and other employing authorities become convinced that the assessment system is valid and reliable and make long-term commitments to support the system (Ingvarson, Kleinhenz & Wilkinson, 2007, p. 17).

Notwithstanding this positive assessment of their potential to change practice, pay for performance schemes may be based on a fundamental misconception: namely that teachers are motivated to change what they do by the prospect of more money.

As in New Zealand currently, the British government developed proposals in the early 2000s to introduce performance management, performance pay and career pathway regulations in England and Wales in order to "modernise" the teaching profession, raise its status and improve student outcomes (Storey, 2000). Indeed the aims, justifications and language of the British proposals are very similar to those being advanced at present by the New Zealand Ministry of Education, and in similar circumstances. At the time, schoolteachers in England and Wales were paid on a salary scale of nine automatic annual step points. Additional points were gained through taking on additional responsibilities or senior management responsibilities.

The Green Paper's modernization proposals included three mandatory individual appraisal objectives: one for improved teaching, one for improved pupil performance targets and one for professional development (p. 511). Appraisal would be required to include classroom observation. In future, high performing teachers would be permitted to receive two step increments annually through a new scale, poor performers none. Teachers who reached the top of the scale and presented a "portfolio of evidence" could apply to cross the threshold and receive a pay increase of approximately ten percent. The portfolio "will include data drawn from classroom observations, analysis of pupils' results, and evidence of commitment to professional development and its impact on classroom performance" (p. 511). It was anticipated that the threshold could be met by a large majority of teachers, after one or more attempts, but that a sizeable minority would not or would not wish to attempt to do so because of the additional commitment required. The Green paper also included a "whole-school performance award in the form of a one-off bonus payment, to between a quarter and third of all schools" (p. 511).

As Storey argues, what is evident from the results of the Government's consultation exercise is that while it clearly assumed that the teacher and headteacher workforces and local education authority employers would be attracted by its "modernization" proposals, a large majority of respondents either were not attracted, or foresaw considerable practical implementation difficulties. Storey comments that such responses are entirely predictable given what is known from performance management literature. First, performance pay "is not a 'Factor X' or an independent variable which, when introduced into a situation, produced a predictable set of effects' (p. 516). It is, rather, one element in a reward system. Second, the design of the system implicitly or explicitly encourages some behaviours and consequences at the expense of others. Storey gives the example of scheme designers anticipating that a certain proportion of teachers will cross the threshold which creates the prospect of a two-tiered teacher service "with some teachers categorized as fully competent and others as in preparatory stages of development or as failures" (p. 516). Naturally, she wonders how parents might react in this situation, presumably in terms of seeking to protect their own child's best interests.

More fundamentally, perhaps, Storey observes that "in essence, a reward system, which introduces performance-based elements is normally operating on the assumption that motivation and behaviour can be motivated through variable pay" (p. 516). This depends on a fairly crude extrinsic view of teachers' motivation - to earn more money - whereas a more appropriate and accurate way of understanding teacher motivation may be "expectancy theory":

> This describes the factors that will influence how much effort an individual is realistically likely to expend. Thus motivation will only be strong if certain conditions are met. In particular, the rewards on offer have to be sufficiently attractive to the individual and be considered commensurate with the amount of effort expended. Second, the individual has to perceive that there is some high probability that the reward will result from the effort expended. There must be some clear perceived link between effort and reward. (p. 517)

In any event the Green Paper's apparent underlying assumptions are not borne out in practice with similar rewards schemes for teachers: evidence is that teacher motivation may be negatively affected and the impact on student achievement may be minimal. Indeed, a more

recent review of studies of public sector performance pay schemes gives considerable cause for concern.

> These studies indicated that the PRP schemes had a limited impact upon employee motivation but that there were a number of negative impacts including deleterious effects on cooperation and team working and divisiveness in the workplace. These were views shared not only by employees but also by their managers. (Farrell & Morris, 2009, p. 80)

Storey (2000) cites authors who suggest that the supposed problems of recruitment, retention and motivation of schoolteachers, used in part to justify the Green Paper, may have nothing at all to do with the absence of a performance-related pay scheme. Similarly, teachers' reported reluctance to adopt one may be due to concerns about technical aspects of the scheme and their awareness that they have often produced relatively poor results in practice (p. 517). Farrell and Morris's survey of 330 teachers during the timeframe for making the first applications to cross the threshold in England and Wales identified four negative responses to the scheme: the (narrow) methods of assessment, the potential for subjectivity and favouritism (headteacher support was a requirement of application), the impact on teacher morale and the bureaucratic burden (2009, p. 89).

As it turned out, Farrell and Morris reported that many teachers in their sample were what they described as "reluctant applicants" - they disagreed with the scheme on principle but believed, for example, that they were underpaid or might be stigmatized if they chose not to apply, or felt peer pressure to apply (p. 89). However, in practice 97 per cent of applicants were successful. This may help partly to explain the findings of a more positive report on the England and Wales scheme by Marsden and Belfield (2006). These authors surveyed teachers and headteachers just before implementation of the scheme, after one year and after four years. Significantly, perhaps, they report that the financial incentives for crossing the threshold proved to be much greater than the ten per cent envisaged in the original Green Paper proposals, "potentially taking teachers" earnings to 25% above the bottom of the lower scale, where some 60% of teachers had been clustered in 2000' (Marsden & Belfield, 2006, p. 4). Moreover, the increments are permanent and count towards superannuation entitlements. The authors identified two distinct school level strategies at work early on, once the scheme got underway. They labelled these "reformer" and "firefighter" strategies.

In the former, “some schools had approached performance management as a means of improving how schools are run, to achieve better coordination between teachers’ activity in the classroom and the school’s wider objectives” (p. 6). In terms of the latter, “the majority of schools were using the new scheme as a means of getting what was felt to be a long overdue pay increase for teachers, and which should ease staff retention problems, especially in high living cost areas such as London” (p. 6).

Four years later, the proportion of “reformer” schools in the sample had reportedly increased. In these schools, the authors report that the successful establishment of performance management had “contributed to improved pupil performance” (p. 29). However, this (statistically significant) improvement was limited to those schools in the sample where the headteacher reported improved “goal setting” as a result of the scheme’s introduction. Among those schools where introduction of the PRP scheme had not been accompanied by improved goal setting, the majority did not improve pupils’ academic performance. The authors cited similar results from another evaluation of the same scheme and concluded that the “statistical results could equally well derive from improved goal setting as opposed to simple financial incentive” (p. 29).

The authors’ argument is that by 2004, despite initial academic scepticism and teacher hostility, performance management had “taken root” (p. 29) in many state schools in England, and was contributing to improved goal setting. In turn the improved goal setting was attributed to “negotiations” between head teachers (or their delegates) and teachers in order to better “align” school and teacher objectives. The fact that head teachers effectively choose which teachers to “put forward” for the threshold assessment and are therefore in a powerful position to “manage” teachers’ willingness to commit to school level goals and priorities is viewed constructively and positively by the authors. Of course another way of viewing the reported effects of the scheme in such schools is that it recasts teacher professionalism as “managed” performance including an increased focus on annual “goal-setting” in line with clear management goals. The extent to which this is a positive modernizing agenda for schooling depends almost entirely on one’s view of the enduring value of individual teacher agency and autonomy.

SUMMARY

This section has shown that:

- The proposal to pay teachers for their "performance" is suggested by research into motivation where it has been observed that paying employees for the results of their efforts can have the effect of motivating improved performance – at least under certain circumstances. One major review of performance pay schemes concluded that: (1) judgments about teacher performance for "high stakes" purposes such as registration, reward, career advancement and promotion must be based on several sources of evidence; (2) performance standards are multidimensional (e.g. knowledge, skills, dispositions), and therefore multiple forms as well as multiple sources of evidence are required. (3) valid and reliable assessment requires independent, trained assessors together with evidence about the context in which judgments about performance are being made.
- Performance pay schemes that are linked with student achievement adopt two quite distinct approaches: (1) teacher performance is based on student performance on standardised tests; or (2) teacher performance is based on evidence about what students are doing in classrooms as a result of conditions for learning established by teachers. In our view teachers may be held responsible only for exercising their professional judgment to create supportive conditions for students to learn: they cannot be held responsible for what, or how much, students learn.
- Another fundamental problem with performance pay schemes is the simplistic assumption that teachers will be positively motivated to change their behaviour because of the prospect of variable pay. However, the effort expended and the rewards need to be commensurate to be positively motivating for the teacher. It is also important that the system does not damage intrinsic motivation by disempowering the teacher. Moreover, required behavioural changes may in fact produce negative effects: for example, on student self-efficacy or classroom environment, or teacher-manager relations, or simply the individual teacher's enjoyment of the tasks.
- Two studies from England claim to have found statistically significant evidence of student achievement gains that were linked with the introduction of a particular performance pay scheme. However, it was conceded by the authors of one of the studies that the scheme's reported achievement gains were just as likely to have been

the result of better goal setting by school managers and teachers, because the majority of schools in which there was no improvement in goal setting showed no improvement in student achievement.

SECTION 8. CONCLUSION

> *Most research articles, after finding a set of things that is correlated with student performance, immediately go to a section on policy conclusions. The steps between the statistical analysis and the section on policy conclusions are seldom discussed." (Hanushek, 1997, p. 303)*

Treasury's policy advice for education is based on a relatively simple economic model of the effects of investing in human capital. The model assumes that variations in student achievement are strongly influenced by variations in teacher quality and can, therefore, be reduced over time by improving teacher effectiveness. This improvement, Treasury argues, can be brought about by providing incentives to teachers based on measures of the achievement of students.

In previous sections we have shown that: (1) teachers are only one of the many factors which influence student learning and achievement: what a child brings to school (ability, culture and values, and previously acquired knowledge and skills) is an even more powerful determinant of what and how much that child will learn; (2) Value Added Measures are entirely unreliable and, therefore, cannot ethically be used as a basis for rewarding teachers. Nevertheless, improving school effectiveness is a worthy goal that we strongly support. If it cannot be achieved by providing incentives for teachers, then how can it be achieved? In this final section we attempt to provide some answers to this question.

In recent years politicians and officials have frequently referred to the work of Sir Michael Barber and Mona Mourshed of the global corporate consultancy McKinsey & Company and, in particular, to their observations about how the supposedly "high performing" systems they identified, succeed.

> High performing school systems consistently do three things well. They get the right people to become teachers (the quality of an education system cannot exceed the quality of its teachers). They develop these people into effective instructors (the only way to improve education is to improve instruction). They put in place systems and targeted support to ensure that every child is able to benefit from excellent instruction

> (the only way for the system to reach the highest performance is to raise the standard of every student). (Barber & Mourshed, 2007, p. 13)

At one level, of course, these observations are simple truisms with which we and most other commentators readily agree. The real policy challenge, however, is *how* does a system recruit the best teaching candidates, how does it ensure that they become successful classroom practitioners and what sorts of systems and targeted support do they put in place for the benefit of all learners? It is at this point that we are reluctant to support the particular policy prescriptions of influential private sector GERM advocates such as the two McKinsey partners. Early on in our report, we stated some other truisms: that schooling is complex, that many teaching and learning variables interact in tacit or implicit ways which are not amenable to direct observation or assessment, and that policy makers and politicians should always proceed cautiously, introducing new policies only where there is evidence both that they will do some good and that they will cause no serious harm. We suggested the possibility that the existing policy settings or "solutions" in the New Zealand context, together with those now being promoted so assiduously by Treasury, may in fact be causing some of the perceived "problems" with our national schooling system. There are other possible education policy settings that need to be debated in light of the available research and it is to these that we turn now in the final section of this report.

1. Is the "Solution" the Problem?

Reference has already been made (see Section 6, Case Study Two) to the work of Gorard who has studied in some detail the high stakes testing regime in England. In a recent paper he reiterates his well-documented conclusion that it is "not yet to possible to identify differentially effective teachers because of the confounding factors and errors in the measurement of pupil progress" (2013, p. 76). However, he points out that most of us (particularly as students or parents) have a strong belief that some teachers are better than others. He draws on other research that he has carried out which indicates that students themselves are sure that there is considerable variation in the quality of the teaching that they experience (p. 76 et seq.). How can we reconcile the research conclusions with common sense? Based on research, he suggests that in England much of the variation is due to: (1) poor teacher selection practices that turn away good candidates in favour of lesser-qualified ones; (2) wide variations in the quality of the institutions that train teachers. A solution

would require “a more coordinated system of application and selection” (p. 78); and (3) the tendency to view teaching narrowly. Instead we should take into account the complexities of teaching. Basically, he argues (from his studies of student reactions to teachers) that it is vital that students learn to trust their teachers, and so trust others in the wider society, and be willing to put themselves out to help others (p. 78). They can learn this only if teachers are “engaged” with students emotionally, treat them decently and, in short, enter into a “relationship of mutual trust and respect” (p. 76).

Thus, it might be argued that the growing tendency of policy makers to focus on narrow measures of academic achievement may itself be contributing to the kind of teaching which students find so alienating and which, therefore, fails to prepare them for life as citizens in a participatory democracy. Accordingly, we may need to find new definitions of “good teaching”.

2. Changes to Education System Policy Settings

There seems to be general popular agreement that teachers make the main in-school difference and, for once, this is borne out by the research. However, those who support GERM tend to argue that this means that system policy settings should emphasise much more supervision and accountability, more explicit skill-training, more precise teaching and learning objectives and achievement targets, and more incentives such as performance pay. In contrast, we argue for: (1) higher trust in teacher collegiality; (2) targeted student engagement and retention policies in the middle and senior secondary years, particularly for students in low decile school communities; and (3) abandoning the emphasis on National Standards and NCEA Level 2 ‘results’.

(1) In New Zealand, GERM advocates tend to regard John Hattie (2009) as their main authority on these matters but while Hattie does emphasise the importance of teachers he does not favour these kinds of policies. He says, for example, “School leaders and teachers need to create school, staffroom, and classroom environments where error is welcome as a learning opportunity, where discarding incorrect knowledge and understandings is welcomed, and where participants can feel safe to learn, re-learn, and explore knowledge and understanding” (p. 239). He goes on to add that what is needed for school improvement is “a caring, supportive staff room, a tolerance for

errors, and for learning from other teachers, a peer culture among teachers of engagement, trust, shared passion, and so on" (p. 240). These are the very attributes which low-trust schemes of accountability inevitably destroy. A reinvigorated system requires more collegial approaches to teaching.

(2) The government has set a target of 85% of 18year olds holding NCEA level 2 or an equivalent qualification by 2017. A small percentage of young people can achieve this after leaving school before the age of 18, but the vast majority will need to do so during secondary schooling.

Overall, 74% of school leavers in 2012 had attained at least NCEA Level 2. The results for subgroups are as follows: girls 78%, boys 71%, Asian students 87%, New Zealand European students 80%, Pasifika students 65%, Māori students 55%, students from decile 1 or 2 secondary schools 58%, and students from decile 9 or 10 secondary schools 90%.

In part, the differences in percentages of students attaining at least NCEA Level 2 while at school are influenced by the differences in how long the different cohorts persist in their schooling. By the end of their year 12 year in 2011, 45% of the Māori students and 48% of the Pasifika students had attained NCEA level 2. One year later, 54% of the Māori students and 64% of the Pasifika students had attained that level. The main factor in the larger gain for Pasifika students (16% compared with 9%) was that a much higher proportion of them stayed in school for the extra year. A similar pattern applied to attaining NCEA Level 1: almost 25% of the Pasifika students who left school with NCEA Level 1 completed that qualification in Year 12 or Year 13.

Reflecting on this information and the government target for NCEA Level 2, Terry Crooks, a leading scholar of educational assessment and for 15 years co-director of the Ministry of Education's National Education Monitoring Project, has suggested that the following issues need to be considered in relation to the government's target (Crooks, personal communication, 2013): First, the percentage of students with such disabilities that attaining the target would never be realistic (perhaps 3%). Second, the additional percentage of students who end Year 6 of primary schooling so poorly prepared that the task of meeting the target is probably too difficult (perhaps 5%-7%).

Third, the additional percentage of students who make insufficient progress during Years 7 to 11 to be realistically equipped to meet the target (perhaps 10%). Crooks believes that these middle school years are a crucial stage in educational persistence, with many students losing faith in the value of schooling or in their own capabilities during these years. Fourth, the additional percentage of students who leave school before Year 13 but who could have attained NCEA Level 2 by staying to Year 13 (perhaps another 10%). Together these groups comprise up to 30% of all school students.

In fact, in 2012 about 30% of school leavers had not attained NCEA Level 2. Assuming that it is realistic at present for about 5% of our young people to attain the equivalent of NCEA Level 2 through work and study by the age of 18 but after leaving school, there is a gap of about 10% to be closed in order to meet the government's 85% NCEA Level 2 target by 2017 (about 75% of 18 year olds in 2012 reached NCEA Level 2 and 74% in 2011).

Crooks notes that a 10% improvement in five years is an ambitious aim, given that it has to be achieved entirely by improvement within the secondary school years or through greater qualification earning between leaving school and the age of 18. He suggests that meeting the target, even within 10 years, would almost certainly require considerable effort in the wider community, to build support for students' educational motivation and persistence among parents and the wider community. That is an important task, but probably requires a longer time scale even than 10 years. The greater persistence of Asian and Pasifika students now evident suggests possibilities to aspire to, and there is a particular challenge with Māori students and students in low decile schools (Crooks, personal communication, 2013).

(3) It follows from Crooks' analysis, that working on National Standards at primary school and on "results" in Years 9-11 at secondary schools is not the way to improve pass rates at NCEA Level 2. *This requires better retention rates at Years 12 and 13*. Creative and inspiring teachers in all subjects may encourage students to stay on in school, which would be better than a narrow focus on skilled performance. The government needs to be aware that too much focus on narrow accountability (i.e. "test results") can operate against attainment by low achieving students: we need to

retain them in school not bore them into leaving.

We have shown in this report that the real relationships between teachers and learners are complex, multiple and many-faceted, unlike the linear, singular and uni-dimensional mathematical models favoured by some economists, academics and bureaucrats. Consequently, to make a difference to real teaching and learning practices, policymakers need to develop a broader appreciation of the quality of teaching and of the multiple complementary policy settings that are needed to encourage and support their improvement.

3. The Message from International Comparative Achievement Studies

Evidence from international studies of student achievement does provide a reasonable basis for questioning the approaches to teacher development and support that are currently in place (that is, the system's "policy settings"). For example, the performance of our primary school students in mathematics and science as tested in Years 5 and 9 in TIMSS suggests that the quality of teaching of science and mathematics may be more or less average compared with that in other participating countries. However, the performance of New Zealand students in science at the end of compulsory schooling as demonstrated by PISA is well above average, indeed it is among the top tier of countries internationally. Assuming that each of the tests is valid and reliable there would appear to be something fairly complex occurring in the primary and secondary schooling years.

Treasury's preferred approach to raising the quality of teaching in these areas is simply to measure the "value added" to pupils' knowledge by the individual teacher and to reward or sanction the teacher accordingly. This is, as we have noted, a business process model of improvement (see Section 1). The underlying assumption is that if New Zealand pupils perform poorly in science and mathematics, the individual classroom teacher must be the sole cause of poor pupil performance. The Minister of Education's reaction to the latest cycle of TIMSS results in science and mathematics has been to reiterate the importance of National Standards implementation; better assessment of these; the production of more robust data for

teachers, trustees and families; and, reportedly, to consider mandatory rote learning of number operations by pupils.[9]

We see the complexities of the challenge somewhat differently and would argue that Government and officials need to consider other, multiple contributory educational causes to this occurrence and to develop multiple contributory policy settings in response. Each of these contributory causes may be relatively insignificant in its own right but together they may explain the effect of relatively poor pupil performance in a particular learning area. This effect may have little or nothing to do with the "value added"' performance of the individual classroom teacher. The key point we want to assert is that educational problems call for broad educational models of improvement.

We know, for example, that unlike many jurisdictions internationally, secondary school pupils in New Zealand are permitted to discontinue taking science after Year 10, and mathematics after Year 11. Arguably, this is because the selection of senior secondary school achievement and unit standards to credit towards NCEA credentials is driven more by personal student preference and the government's desire to maximize the proportion of students who reach NCEA Levels 1-3, than it is by the priority to provide a broad and balanced educational experience for all pupils as, for example, in the international baccalaureate system.

We know also that early childhood and primary initial teacher education programmes tend not to attract the academically highest performing secondary school students nor necessarily those with broad and balanced depth of senior secondary school study. On this basis it seems reasonable to suggest that a sizeable proportion of candidates for initial teacher education (ITE) in the primary school sector may not have sufficient understanding of either mathematics or science. The logical consequence of this situation is that New Zealand primary schools are likely to have too few specialist teachers of science and mathematics as well as insufficient science and mathematics knowledge and skills across all their teachers.

Until relatively recently, under-preparation at secondary school may not have mattered greatly. Historically, the teachers' colleges offered primary teacher education candidates a

[9] http://beehive.govt.nz/release/international-studies-show-changes-must-be-made-improve-education-outcomes; http://www.odt.co.nz/news/national/246954/parata-pondering-return-basic-arithmetic-schools

broadly based, participatory liberal arts curriculum which included opportunities for them to develop greater knowledge, skills and understanding of mathematics and science (and the teaching of these) than is currently the case. However, Government funding for primary initial teacher education has declined approximately 20 percent in real terms since 1993 (O'Neill, 2012). At the same time, student fees have increased to attempt to fill the gap and many have had to combine substantial part-time work with full-time study in order to make ends meet. These students may have limited opportunities therefore to immerse themselves in the co-curricular and field-based authentic learning opportunities that traditionally were regarded as valuable formative experiences for prospective primary teachers. In this context, then, the present government's plan (Education Review, 2012) to make primary school initial teacher education programmes postgraduate (one year) entry only is of serious concern. Indeed, to our knowledge, there has been no consideration in any policy announcements to date of the broad subject knowledge and pedagogy needs of primary school teachers, and how these may be assured in the new model.

Finally, add to this the contemporary education policy setting's dominant emphasis on functional literacy and numeracy, on National Standards and on ensuring that pupils work towards and meet these standards from age 5. Our approach to the issue of teacher quality or performance would suggest that each of these conditions (the structure of NCEA, secondary school curriculum and career choices, ITE funding, changes in the ITE curriculum, National Standards) is very likely a contributory cause to what may be a comparatively poor quality of science and mathematics teaching in primary schools. *If this is true, simply measuring, rewarding and sanctioning individual teachers will do nothing to address or remedy the problem.*

Against this, the Ministry of Education might point to the considerable system-level efforts that have been made since the late 1990s to improve the quality of numeracy teaching nationally through the provision of face-to-face and online professional development programmes and associated teaching resources.[10] However good and extensive this professional development and material support for teachers may be, though, if the intended recipients begin with a poor or limited understanding of mathematics, it may not produce, or be capable of producing, the desired effects of improved numeracy teaching and learning.

[10] http://www.nzmaths.co.nz/what-numeracy-project

Taking a broader view, then, we argue that to address the "problem" of relatively poor pupil achievement in mathematics, government needs to consider not just teacher performance but teacher knowledge and skills, specifically teachers' understanding of mathematics. On this broader view, the quality of the pool of potential candidates for teacher education is of importance and this, in turn, relies on the breadth and depth of secondary schooling. If generalist primary school teachers are to teach all the learning areas of the primary school curriculum, they must have breadth and depth of study in these areas. If ITE providers are to support students to develop the range of pedagogical and assessment strategies necessary to diagnose pupils' understandings and misunderstandings, and to respond to these confidently, then ITE must be adequately funded and of sufficient duration to undertake this task. Similarly, ongoing professional development for teachers to be able to use evidence-based curricula such as the Numeracy Project (as opposed to the non-evidence-based rote learning proposition of the Minister) would appear to be highly desirable. The Numeracy Project emphasises understanding of strategies but facility with their application is also important and is often being neglected. Similarly, the more recent Science Learning Hub provides very engaging curriculum resources for teachers of science to use in Years 2 to 10, but offers no way to increase their confidence to teach science, or their science knowledge, or to expand their science pedagogical repertoire.[11]

With respect to science in the primary school, the Prime Minister's Chief Science Advisor (Gluckman, 2011, pp. B-61-B67) has produced a complex analysis and set of recommendations that are consistent with our assessment of the real issues to be addressed. According to his report, New Zealand primary school pupils spend less time engaged in science than overseas pupils, while primary trained teachers have less pre-service specialization in science teaching, and less ongoing professional development than overseas teachers. Professor Sir Peter Gluckman advocates pre-service courses in science and its pedagogy (precisely the courses that we claim have been discontinued in primary ITE since 1993 because of government funding cuts) on the basis that, "The tendency of trainees to lack a background in science indicates the need for foundation courses that develop understanding of the history and philosophy of science and core concepts. The second area

[11] http://www.sciencelearn.org.nz/

for development is the understanding of pedagogical methodologies that support effective integration of science into primary teaching and learning programmes" (p. B-62).

Notably, Gluckman also identifies issues with the teaching of science in the secondary school, for example the need to cater for pupils who wish to pursue science careers and for those who do not but nevertheless need to develop an understanding of and enthusiasm for science as it relates to their lifeworlds. Significantly, the strategies that Gluckman advocates depend for their success on a national co-ordinated effort, which creates and disseminates communally owned science teaching knowledge (content and pedagogy) across groups, schools and the system as a whole. It is a collaborative model, not a competitive one of the sort that Treasury's approach to teacher quality and performance management would undoubtedly foster. Sadly, as Cathy Wylie has observed (2012), one of the unfortunate consequences of the Tomorrow's Schools reforms has been to remove the "vital connections" that bind an organic education system together and actively encourage the sharing of good teaching and learning practice.

4. Changes to Teacher and School Improvement Policies

It is clear that one of the goals of the present government is to make schools more 'effective' than they are. We believe that in this report we have shown that international studies reveal that the New Zealand school system as a whole ranks highly with respect to student achievement when compared with the school systems of other countries and states with similar cultures. This means that the government's policy is a high-risk one since changes to complex and well-functioning systems often have the effect of degrading rather than improving performance. We believe that changes should be made only when: (i) there is sound evidence of positive effects or the idea has been carefully piloted; and (ii) there is minimal risk of systemic damage. We believe that we have shown clearly that VAM approaches to teacher evaluation and performance pay fail this test: the international evidence is not supportive and the damage to the profession and children's learning may be substantial.

In this section we argue that six policies, if carefully implemented, have the potential to make New Zealand schools even more effective than they are already. These six policies are as follows:

1. Policies which increase the level of teaching skill of all classroom teachers in the system, Increased student achievement is only possible if students receive more effective teaching. A recurring reason for school reform failures is "because they assumed that teachers would know how to do things they actually didn't know how to do" (Schleicher, 2011, p. 17). A striking example of a New Zealand reform which failed in this way was the decision at the end of the 1990s to move to a policy of full inclusion without first providing all classroom teachers with the additional skills which were going to be needed in order to teach the children with various kinds of disability who were about to be mainstreamed. To acquire a new skill it is necessary to engage in active rehearsal of that skill with feedback (provided, for example, by another teacher) and to continue the practice-feedback cycle until the new skill has been mastered (Timperley, Wilson, Barrar & Fung, 2007). The only way to develop more advanced teaching skills system wide is to commit the necessary resources to improved pre-service and in-service education.

2. Policies which give first priority during pre-service and in-service education to those diagnostic, teaching and evaluation skills which have been shown to be the most effective in fostering different kinds of learning. Increased levels of teaching skill are only likely to occur if everyone in the education system starts giving priority to the dissemination of those diagnostic, teaching and outcome monitoring skills that have been shown by research to be more effective in fostering learning than those in current use. By way of example, Tunmer et al (2013) argue that the National Literacy Strategy adopted by the NZ Ministry of Education some 15 years ago has failed because the Ministry's policy ignored the results of the scientific research into the teaching of reading to disadvantaged students. Sound research into what works in the classroom has been accumulating for 30 years and provides a rich resource for any school system which wants to improve the effectiveness of its classroom teaching.

3. Policies which support improved teaching by increasing the availability to teachers of teaching materials which have been field-tested and shown to be effective in developing each of the many learning outcomes contained in the New Zealand school curriculum. The New Zealand Ministry of Education (and its predecessor the Department of Education) have a proud history of educational publishing. In general the Ministry and its school publications branch have focused on providing resources

for teachers to use – leaving teachers free to decide how these resources will be used in the classroom. However, increased levels of student achievement are more likely if classroom teaching is supported by the progressive development of teaching materials which have been field tested and revised until they have been shown to be effective in developing important learning outcomes.

4. Policies that upgrade the management skills, appraisal skills and professional development skills of all in the system. School leaders play a central role in developing the teaching skills of classroom teachers. The introduction in New Zealand of the evidence based programmes contained in the Positive Behaviour for Learning (PB4L) initiative occurred because system leaders in the Special Education division of the Ministry of Education recognised and understood the importance of moving from practices which merely sound plausible to practices which have been shown, in controlled evaluations, to be demonstrably effective. School leaders play a central role in organising professional development and in motivating continued improvement on the part of all staff within their schools.

5. Policies that achieve system-wide improvements in the level of understanding of sound research methods in education and their results. Decisions will be most reliable if they are based on the results of sound empirical research into the relative effectiveness of different ways of teaching different kinds of skills and understandings to students at different levels of social and academic development. This means that improvements in teaching effectiveness will be closely dependent upon the speed with which (a) teachers come to understand what counts as trustworthy research, (b) teachers begin to select pre-service and in-service education programmes which prioritise the learning of evidence based teaching practices, and (c) teachers themselves begin to use the results of research into what works as they go about the task of deciding what to teach and when and how to teach it.

6. Policies that are designed to encourage collaboration and whole school improvement by capitalizing on the intrinsic motivation of teachers and school leaders. One of the characteristics of teaching as a profession is that the primary motivation of most teachers is intrinsic. Intrinsic motivation is motivation that is sustained by the

knowledge that one is doing an important job well and can see the results in student development.

Ill-considered change has the effect of reducing intrinsic motivation and creating resistance. Trust cannot be legislated for. A recurring reason for school reform failures is "because teachers and schools did not buy in to the reform strategy" (Schleicher, 2011, p. 17). It follows that system wide improvements in student achievement are likely to be achieved only by policies that capitalize on teachers' intrinsic motivation rather than by policies which result in teachers competing for extrinsic rewards.

In order to maintain this intrinsic motivation, it is essential that teachers be involved as equal partners in the design and implementation of reform. School reform will not work unless it is supported from the bottom up. In looking for examples of overseas jurisdictions where most or all of these six improvement policies may be evidenced, we are heartened by the story of Finland. Sahlberg (2007) draws on the Finnish experience of educational reforms since the 1970s and argues that what is absent from the orthodox reform agenda (GERM) is an acknowledgement of the importance of the applied and social aspects of learning that mirror everyday activities. He advocates "flexibility and loose standards" as opposed to standardisation, "broad learning combined with creativity" instead of an undue emphasis on literacy and numeracy, and "intelligent accountability with trust-based professionalism" (p. 152). All of these are utterly consistent with our suggestions as to how to improve our system.

Proponents of New Zealand's "reforms" may seek to dismiss this view as outdated liberal-progressive ideology, but the fact is that it works both educationally and in terms of equity. Finland has one of the most successful and equitable education systems in the world (OECD, 2008, 2011). Michael Fullan, an international expert on school reform, has compared policies which produce system-wide improvement with those which do not. According to him, policies work if they: (a) foster the intrinsic motivation of teachers and students; (b) engage teachers and students in the continuous improvement of instruction and learning; (c) inspire collective team work; and (d) affect all teachers and students. These are effective because they work on changing the culture of the system, operate on both practices and relationships and lead to whole system improvement (Fullan, 2011, p. 5).

In conclusion, we return to our earlier analysis:

1. According to the OECD report (2005) “teaching is a profession in long-term decline. As societies have become wealthier and educational qualifications have increased and employment opportunities have expanded, teaching’s appeal…does seem to have diminished” (p.5).
2. However, there is increased awareness of how crucial teachers can be in the achievement of students and in the progress of society.
3. Yet, many societies, including our own, seem bent on instituting measures (such as constant surveillance and performance management) which alienate teachers, deter students (especially the most gifted) from a teaching career, encourage unprofessional behaviour (such as cheating on tests) and lead to rapid turnover of teachers in a system that needs stability.

We argue that alternative policies are required to enhance and upgrade the status of the teaching profession, to attract and retain high achieving candidates and to allow them to flourish as autonomous professionals. Although we do not claim to have all the answers, we submit that the evidence we have presented strongly suggests that the agenda we put forward is more defensible than that of the New Zealand Treasury, and we welcome informed debate on it.

APPENDIX: TECHNICAL ISSUES

Variance and Correlation

The term "variance" is used frequently in educational research with three common understanding of its meaning. Technically, the variance is the expected value of the average squared deviation from the mean and is the square of "the standard deviation" for that variable. Put more simply, the variance is a measure of the spread of a set of data points around their mean value.

A second common usage of the term variance is in relation to the correlation between two variables. Put simply, if we plot a graph of two variables then the correlation coefficient is a measure of how closely the data points match a theoretical line or curve. While the correlation coefficient describes the relationship, the square of the correlation coefficient indicates the proportion of the variance on one variable which is linearly associated with the varying scores on the other correlated variable.

For example the term "variance" is used in phrases such as: "Our results are largely driven by findings from the literature and new analyses that more than 90 percent of the variation in student gain scores is due to the variation in student-level factors that are not under the control of the teacher (Schochet & Chiang, 2010, p.35). While it may seem reasonable in this instance, we need to be cautious before reading too much causality into this statement. We do not know whether they are both "dependent" on changes in a third, or greater number of, variable(s). The existence of a mathematical correlation, however, is often taken as evidence of a causal relationship where one may not exist. This may be because the terms that are often used to indicate the degree of this relationship such as "explained", "due to" or "attributable" which incorrectly imply causality. Any causality will need to be established separately.

A third, but less common, use of the term "variance" is as a layperson's term to describe a general measure of the spread of a set of data. Often used in the media and other less formal documents, the use of the term in this way is designed to give the reader a sense of the distribution of the data. Some common colloquial uses of the term "variance" include the

"range" of the data (the difference between the smallest and largest value), the "interquartile range" (the difference between the lower and upper quartile), and the range between the 5^{th} percentile and the 95^{th} percentile, commonly used in international assessment projects such as PISA.

Controlling for Other Variables

When carrying out scientific work the aim is often to discover the strength of some "variable". Thus for example, a medical researcher might want to find out if losing weight can lower blood-sugar levels. In order to do that, the researcher might take a group of subjects with diabetes, encourage them to lose a certain amount of weight and then measure any change in blood sugar readings. If there is a strong correlation between weight loss and lower blood sugar, it would be tempting to say that the study has shown that weight loss leads to lower levels of blood sugars. But this would be premature: the researcher would need to rule out (in technical terms, "control for") other variables such as the age or gender of the subjects: dieting might reduce sugar levels in young people but not in old people, in women but not in men. Similarly, occupation, diet, home situation and exercise levels might all make some difference to the results. It is therefore not obvious that researchers can in fact control adequately: all variables occur in a context and often interact with each other. The average reader of research needs to be properly sceptical as to how well such "controls" work (even in medicine).

It is particularly difficult to control for variables in educational research. Here (as in much social science research), a major challenge is to control for SES (Socioeconomic Status) and researchers almost invariably state that they have controlled for SES in the same way as medical researchers try to control for gender and diet. As we have seen, it is not easy to control for health factors. It is even more difficult to use a statistical device to capture the effects of SES. In New Zealand, it is common to use the decile level of the school as the mechanism for controlling for SES. The decile level of a school (useful as it is for funding purposes) is the result of a complicated formula which can obscure much of what is going on. A small high decile school in a rural area might be very different in its composition from a large one in a city, while a low decile school with students drawn from a tightly knit Pasifika community with a shared culture and a shared faith might be very different from one with a more diverse multi-cultural enrolment. School rankings such as those produced

annually by Metro do not acknowledge this. In short, it is far from clear whether social science and education researchers can in fact control adequately for all variables since they occur in a complex context and often interact with each other. This need to be kept in mind at all time when looking at large and complex studies of schooling which make claims to have "carefully controlled" for all variables other than the quality of the teacher.

Effect Sizes

Effect sizes are a way of describing the strength of the difference in the performance of two groups on the same outcome measure (e.g. a test of mathematical knowledge). The difference in mean outcome score of the two groups is then divided by the standard deviation of the underlying population to get a standardised measure of difference (In practice, the standard deviation of the population is rarely known so it must be estimated either from the standard deviation of the "control" group, or from a "pooled" value from both groups). This is the effect size. For example, if there is a difference of 8 points between two sets of IQ data (SD =15) we divide 8 by 15 to get an effect size of 0.53.

Effect sizes are often used in meta-analyses of a cluster of research studies to try to examine the extent to which their results are similar and how strong the effects are. Thus, the results of studies can be plotted along a continuum from large negative effect size to very large positive effect size. In trying to assess whether a difference is note-worthy, a cut-off point has to be selected. This is a hazardous exercise, as it may mean that potentially important effects may be overlooked or the importance of a small effect size be exaggerated. In his synthesis of numerous meta-analyses, John Hattie required effect sizes of at least 0.4 for an effect to be regarded as important (Hattie, 2009). A danger with such approaches is that studies of widely varying quality and size can count equally in the resulting effect size. (See Snook et al, 2009).

A writer on medical research provides a caution: "Although quantitative methods such as the computation of effect size play a crucial role in evidence based medicine it will never fully replace the evidence collected by informed physicians seeking to optimize the care of their patients" (Faraone, undated, p. 4). The significance of this for education cannot be ignored. If professional interpretation is needed by a doctor face to face with one patient for twenty minutes and probably focused on one problem (e.g. her high Blood Pressure), how

much more so is this required in a classroom where a teacher has to deal for a whole day with 30 or more active children (or in the case of the secondary school, perhaps up to 200 adolescents) with many different problems.

REFERENCES

Baker, E.L., Barker, P.E., Darling-Hammond, L., Haertel, E., Laddl, H., Linn, R., Ravitch, D., Rothstein, R., Shavelston, R., & Shepard, L. (2010). *Problems with the use of student test scores to evaluate teachers*. Briefing Paper. Economic Policy Institute 27B, August 29.

Barber, M. & Mourshed, M. (2007) *How the best performing school systems come out on top.* London: McKinsey & Co.

Benn, M. & Millar, F. (2006). *A comprehensive future: Quality and equality for all our children*. London: Compass.

Biddulph, F., Biddulph, J., & Biddulph, C. (2003, June). *The complexity of community and family influences on children's achievement in New Zealand: Best Evidence Synthesis Iteration*. Wellington: Ministry of Education.

Birks, S., Snook, I., Prochnow, J., Rawlins, P., & O'Neill, J. (2013). Econometrics for educational policy: the use of value-added measures of teacher performance. Massey University Centre for Public Policy Evaluation. Issues Paper No. 16. Available: http://cppe.massey.ac.nz/papers/cppeip16/cppeip16.pdf

Borman, G. D., & Kimball, S. M. (2005). Teacher quality and educational equality: Do teachers with higher standards-based evaluation ratings close student achievement gaps? *The Elementary School Journal, 106*(1), 3-20.

Bray, M. (2009). *Confronting the shadow education system: What government policies for what private tutoring?* Paris: International Institute for Educational Planning, UNESCO. Retrieved from: www.iiep.unesco.org/.../pdf/2009/Bray_Shadow_education.pdf

Burris, C. & Welner, K. (2011). *NEPC policy memo: Letter to Secretary of Education Arne Duncan concerning evaluation of teachers and principals*. Boulder, CO: National Education Policy Center. Retrieved from: http://nepc.colorado.edu/publication/letter-to-Arne-Duncan

Cervini, R. A. (2008). Class, school, municipal and state effects on mathematics achievement in Argentina: A multilevel analysis. *School Effectiveness and School Improvement, 20*, 319-340.

Charter schools: Additional Federal attention needed to help protect access for students with disabilities. (2012). Retrieved from: http://www.gao.gov/assets/600/591435.pdf

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2012). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. Retrieved from: http://www.nber.org/papers/w17699

Coyne, M. D., Kameenui, E. J. & Carnine, D, (2007). *Effective teaching strategies which accommodate diverse learners* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Crooks, T. 2013). Personal communication with the authors.

Cruickshank, D. R. & Heefele, D. L. (1991). Research-based indicators: Is the glass half-full or half empty? *Journal of Personnel Evaluation in Education, 4*, 33-39.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching.* Alexandria VA: Association for Supervision and Curriculum Development.

Dee, T., & Keys, B. (2004). Does merit pay reward good teachers? Evidence from a randomized experiment. *Journal of Policy Analysis and Management, 23*(3), 471-488.

Education Counts (Ministry of Education). (2012). *Annual expenditure per student*. Retrieved from http://www.educationcounts.govt.nz/__data/assets/pdf_file/0003/15726/Indicator-inID-2043.pdf

Education Review (2012). *Failure to launch: postgraduate initial teacher education*. Downloaded from: http://www.educationreview.co.nz/education-in-review/december-2012/failure-to-launch-postgraduate-initial-teacher-education/#.UhlfJc0lU_p

Education Review Office (2005). *Annual report for the year ended 30 June 2005.* Wellington: Education Review Office.

Education Review Office (2012). *Science in the New Zealand curriculum: Years 5-8.* Wellington: Downloaded from http://www.ero.govt.nz/National-Reports/Science-in-The-New-Zealand-Curriculum-Years-5-to-8-May-2012

Faroane, S. (undated). Understanding effect size: How it's measured and what it means. Retrieved from: www.medscape.org

Farrell, C. & Morris, J. (2009). Still searching for the evidence? Evidence-based policy, performance pay and teachers. *Journal of Industrial Relations, 51*(1), 75-94.

Fea, S. (2013, 1 July). Minister: I don't like deciles. *The Southland Times*. Retrieved from http://www.stuff.co.nz/southland-times/news/8860387/Minister-I-don-t-like-deciles

Fergusson, D.M, Horwoood, L.J, & Boden, J.M. (2008). The transmission of social inequality: Examination of the linkages between family socio-economic status in

childhood and educational achievement in young adulthood. *Research in Social Stratification and Mobility, 26,* 277-295.

Fullan, M. (2011). *Choosing the wrong drivers for whole system reform.* Centre for Strategic Education, Seminar series paper No. 204.

Glass, G. (2012).The Problem with Value-Added Measures. National Education Policy Center. Retrieved from: http//nepc.Colorado.edu/blog/problem-value-added-measures

Gluckman, P. (2011). *Looking ahead: Science education for the twenty-first century*. Auckland: Office of the Prime Minister's Science Advisory Committee.

Goe, L., Bell, C., & Little, O. (2008). Approaches to evaluating teacher effectiveness: A research synthesis. Washington, DC: National Comprehensive Center for Teacher Quality.

Gorard, S. (2006). Value-added is of little value. *Journal of Educational Policy, 21* (2), 233-241.

Gorard, S. (2008). The valued added of primary schools: what is it really measuring? *Educational Review 60*(2), 179-185.

Gorard, S. (2012). Now you see it, now you don't: school effectiveness as conjuring? *Research in Education 86,* 39-45.

Gorard, S. (2013). What difference do teachers make? A consideration of the wider outcomes of schooling. *Irish Educational Studies, 32*(1), 69-82.

Gray, J., Jesson, D., & Jones, B. (1986). Towards a framework for interpreting examination results, in R. Rogers (ed.), *Education and Social Class,* London: Falmer Press.

Hanushek, E. A. (1997). Outcomes, incentives, and beliefs: Reflections on analysis of the economics of schools. *Educational Evaluation and Policy Analysis, 19*(4), 301-308.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review, 30*(3), 466-479.

Harker, R. (1995). Further comment on "Do Schools Matter?". *New Zealand Journal of Educational Studies, 30* (1), 73-76.

Harker, R., & Tymms, P. (2004). The effects of student composition on school outcomes. *School Effectiveness and School Improvement, 15,* 177–199.

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

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

House of Representatives (2008). Inquiry into making the school system work for every child. Report of the Education and Science Committee.

Ingvarson, L., Kleinhenz, E., & Wilkinson, J. (2007). *Research on performance pay for teachers*. retrieved from: http://research.acer.edu.au/workforce/1

Jacob, B. A., Lefgren, L., & Sims, D. (2008). The Persistence of Teacher-Induced Learning Gains Available from http://www.nber.org/papers/w14065

Kyriakides, L., & Creemers, B.P.M. (2008). Using a multidimensional approach to measure the impact of classroom-level factors upon student achievement: A study testing the validity of the dynamic model. *School Effectiveness and School Improvement, 19*, 183-205.

Lavy, V. (2002). Evaluating the effect of teachers' group performance incentives on pupil achievement. *Journal of Political Economy, 110*(6), 1286-1317.

Leigh, A. (2010). Estimating teacher effectiveness from two-year changes in students' test scores. *Economics of Education Review, 29*(3), 480-488.

Lemke, M., Lippman, L., Bairu, G., Calsyn, C., Kruger, T., Jocelyn, L., Kastberg, D., Liu, Y.Y., Roey, S. & Williams, T. (2002). *Outcomes of learning: Results from the 2000 Program for International Student Assessment of 15-year-olds in Reading, Mathematics and Science Literacy*. Washington: US Office of Education. ED468749

Lucas, R. E. (1976). Econometric policy evaluation: A critique. In K. Brunner & A. H. Meltzer (Eds.), *The Phillips curve and labor markets* (pp. 19-46). North-Holland: Elsevier B.V.

Luyten, H. (2003). The size of school effects compared to teacher effects: An overview of the research literature. *School Effectiveness and School Improvement, 14*, 31-51.

Marsden, D., & Belfield, R. (2006). Pay for performance where output is hard to measure: The case of performance pay for school teachers. In D. Lewin & B. Kaufman (Eds.), *Advances in Industrial & Labor Relations, 15*, 1-34.

Medley, D. M., & Coker, H. (1987). How valid are principals' judgments of teacher effectiveness? *The Phi Delta Kappan, 69*, 138-140.

Milankowski, A. (2004). The relationship between teacher performance evaluation score and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79*(4), 33-53.

Ministry of Education. (2007). *NEMP Writing Results 2006*. (41) Wellington: Ministry of Education.

Ministry of Education. (2008). *NEMP Science Assessment Results 2007*. (44). Wellington: Ministry of Education.

Ministry of Education. (2009a). *NEMP Reading and Speaking Assessment Results, 2008*. (49). Wellington: Ministry of Education.

Ministry of Education, (2009b). National Standards and Reporting to Parents. Wellington: NZ Government.

Ministry of Education. (2010a). *PISA 2009: Our 21st century learners at age 15*. Wellington: Ministry of Education.

Ministry of Education. (2010b). *NEMP Mathematics Assessment Results* 2009 (52). Wellington: Ministry of Education.

Ministry of Education. (2011). *NEMP Writing, Reading and Mathematics Report 2010*. (53). Wellington: Ministry of Education.

Ministry of Education (2012). *Career pathways for teachers*. Draft discussion document for internal circulation only. Released under the Official Information Act.

Ministry of Education. (2013). *Key findings from New Zealand's participation in the Progress in International Reading Literacy Study (PIRLS) and Trends in International Mathematics and Science Study (TIMSS) in 2010/11*. Wellington: Ministry of Education

Ministry of Education (undated) Success for all – Every school, every child. Retrieved from: http://www.minedu.govt.nz/NZEducation/EducationPolicies/SpecialEducation/OurWorkProgramme/SuccessForAll.aspx

Mourshed, M., Chijioke, C. & Barber, M. (2010) *How the world's most improved school systems keep getting better*. Retrieved from: http://mckinseyonsociety.com/how-the-worlds-most-improved-school-systems-keep-gettingbetter/

Nuthall, G. (2002). Knowing what we know and what we need to know about effective teaching. Paper presented to the New Zealand Council for Educational Research symposium, Auckand, October.

Nuthall, G (2007). *The hidden lives of learners*. Wellington: NZCER Press.

OECD (2005). *Teachers matter: Attracting, developing and retaining effective teachers*. Overview. Paris: OECD. Retrieved from: http://www.oecd.org/dataoecd/39/47/34990905.pdf

OECD (2008). *Growing unequal? Income distribution and poverty in OECD countries*. Retrieved

from:http://www.oecd.org/social/soc/growingunequalincomedistributionandpovertyinoecdcountries.htm

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

OECD (2010b). *The high cost of low educational performance: The long-run economic impact of improving PISA outcomes*. Paris: OECD Publishing. doi: 10.1787/9789264077485-en

O'Neill, J. (2012). Rationality and rationalisation in teacher education policy discourse in New Zealand. *Educational Research, (54)*2, 225-237.

Opdenakker, M. & Van Damme, J. (2000). The importance of identifying levels in multilevel analysis: An illustration of the effects of ignoring the top or intermediate levels in school effectiveness research. *School Effectiveness and School Improvement, 11*, 103–130.

Palardy, G. J. (2007). Differential school effects among low, middle, and high social class composition schools: A multiple group, multilevel latent growth curve analysis. *School Effectiveness and School Improvement, 19*, 21-49.

Park, J. (2013). Detoxifying school accountability: The case for multi-perspective inspection. London: Magdalen House (DEMOS).

Peetsma, T., van der Veen, I., Koopman, P., & van Schooten, E. (2006). Class composition influences on pupils' cognitive development. *School Effectiveness and School Improvement, 17*, 275-302.

Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Rothstein, J. (2011). *Review of 'Learning about Teaching'*. Boulder, CO: National Education Policy Center. Retrieved from: http://nepc.colorado.edu/thinktank/review-learning-about-teaching

Rowan, B.R., Correnti, R., & Miller, R.J. (2002). What large scale research tells us about teacher effects on student achievement: insights from the prospective study of elementary schools. *Teachers College Record, 104, 1525-1567*

Sahlberg, P. (2007). Education policies for raising student learning. *Journal of Education Policy*, *22*(2), 147-171.

Sahlberg, P. (2011). *Finnish lessons: What can the world learn from educational change in Finland?* New York: Teachers College Press.

Schleicher, A. (2011). *Building a high-quality teaching profession: Lessons from around the world.* Paris: OECD Publishing.

Schochet, P.N. & Chiang, H.S. (2010). Error rates in measuring teacher and school performance based on student test score gains. Retrieved from: http://ies.ed.gov/ncee/pubs/20104004/pdf/20104004.pdf

Sloane, F., Oloff-Lewis, J., & Hee Kim, S. (2013). Value-added models of teacher and school effectiveness in Ireland: Wise or otherwise? *Irish Educational Studies, 32*(1), 37-67.

Snook, I., Clark., J., Harker R., O'Neill A-M., O'Neill, J. (2009). Invisible learnings: A commentary on John Hattie's book Visible Learning: a Synthesis of over 800 Meta-analyses Relating to Achievement. *New Zealand Journal of Educational Studies, 44*(1), pp. 93-106.

Snook, I. & O'Neill, J (2010). Social class and educational achievement: beyond ideology. *New Zealand Journal of Educational Studies, 45*(2), 3-18.

Snook, I., O'Neill, J., Boyack, J., Brown, S., Clark, J., Greaney, K., Harker, R., Openshaw, R., Prochnow, J., Tunmer, W., Vossler, K. & Walsaw, M. (2012). *Charter Schools for New Zealand: An investigation designed to further the debate in New Zealand on education policy in general and on charter schooling in particular.* Palmerston North: Massey University College of Education.

Storey, A. (2000). A leap of faith? Performance pay for teachers. *Journal of Education Policy, 15*(5), 509-523.

Stringer, E. T. (2007). *Action research* (3rd ed.). Los Angeles: Sage Publications.

Taylor, E.S. & Tyler, J.H. (2011).The effect of evaluation in performance. Evidence from longitudinal student achievement data of mid-career teachers. NBER Working Paper 1687. Retrieved from: http://www.nber.org/papers/w16877

Teddlie, C., Reynolds, D., & Sammons, P. (2000). The methodology and scientific properties of school effectiveness research. In C. Teddlie & D. Reynolds (Eds.), *The international handbook of school effectiveness research* (pp. 55 -133). London: Falmer Press.

Teddlie, C. & Reynolds, D. (2001). Countering the critics: Responses to recent criticisms of school effectiveness research. *School Effectiveness and School Improvement, 12*, 41-82.

Timperley, H., Wilson, A., Barrar, H. & Fung, I. (2007). *Teacher professional learning and development.* Wellington: Ministry of Education.

Treasury (2009). *Challenges and choices: New Zealand's long-term fiscal statement*. Wellington: Downloadedfrom: http://www.treasury.govt.nz/government/longterm/fiscalposition/2009/16.htm

Treasury (2011a) *Treasury report: Strengthening the education workforce*. Report No. T2011/2112. Wellington: Author. Released under the Official information Act.

Treasury (2011b). *Compulsory schooling: possible future directions*. Document partly released under the Official Information Act. Retrieved from: http://www.treasury.govt.nz/publications/informationreleases/education/partnershipschools/index.htm

Treasury (2012). *Treasury's Advice on Lifting Student Achievement in New Zealand: Evidence Brief*. Retrieved from: http://www.treasury.govt.nz/publications/media-speeches/speeches/economicleadership/sanz-evidence-mar12.pdf

Tunmer, W. E., Chapman, J. W., Greaney, K. T., Prochnow, J. E., & Arrow, A. W. (2013). Why the New Zealand National Literacy Strategy has failed and what can be done about it: Evidence from the Progress in International Reading Literacy Study (PIRLS). Palmerston North: Massey University Institute of Education.

Woulfe , C. (2013). All shall pass. *New Zealand Listener,* May 11-17, pp.17-25.

Wu, M. (2012). Abuse of statistics. Linking student test scores to teacher accountability. Retrieved from: http://www.etaq.org.au/wp-content/uploads//TeacherPerformanceAbusersOfStatistics.pdf

Wylie, C. (2012). *Vital connections: Why we need more than self-managing schools*. Wellington: NZCER.

Members of the Educational Policy Response Group for this Project

Ivan Snook is Emeritus Professor of Education, Massey University
John O'Neill is Professor of Teacher Education, Massey University Institute of Education
Stuart Birks is Senior Lecturer, Director of the Centre for Public Policy Evaluation, Massey University
John Church is Adjunct Senior Fellow, Department of Psychology, University of Canterbury
Peter Rawlins is Senior Lecturer, Massey University Institute of Education

Acknowledgment

The authors acknowledge the significant contribution of Emeritus Professor Terry Crooks to successive drafts of this paper. His detailed comments have improved the document greatly. He is, of course, in no way responsible for the opinions expressed or for any mistakes in citation or interpretation which remain.

About the Education Policy Response Group

Massey University's Education Policy Response Group (EPRG) has a history dating back to the early 1980s. It is an ad hoc, eclectic grouping of policy scholars and other interested staff and postgraduate students at Massey University. Its purpose is to meet when necessary to prepare considered responses to major government education policy initiatives and other topical educational debates. The EPRG's first venture was an analysis of the proposed Core Curriculum issued by then Minister of Education, Hon Merv Wellington, in 1983. Since then it has issued reports on many topics such as the curriculum, teacher education and student assessment. Its most recent reports were a critical appraisal (Snook et al, 2009) of Professor John Hattie's 2009 book, *Visible Learning: A Synthesis of Over 800 Meta-analyses Relating to Achievement* and an examination (Snook et al, 2012) of the Charter Schools proposal: *Charter Schools for New Zealand: An investigation designed to further the debate in New Zealand on education policy in general and on charter schooling in particular*.

Institute of Education | Massey University Manawatu | Private Bag 11-222 | Palmerston North 4442 | New Zealand