BUILDING EVALUATIVE CAPABILITY IN SCHOOLING IMPROVEMENT

POSITION PAPER 3

Analysing Student Achievement Data for Measuring Progress in Schooling Improvement

Rachel Dingle

Introduction

This paper presents an overview of how student achievement data and other student level data collected on valued outcomes can be used to create high-quality evidence to inform schooling improvement. The evidence should form the basis for inquiry to understand what needs to change to make a difference to student learning and later to check whether everyone's improvement efforts have been successful.

Schooling improvement involves activities at Ministry, cluster, school and classroom levels. This paper concentrates on the part of a schooling improvement initiative that is centred on analysing and using student data at the school and cluster levels.

For each of the concepts explored we outline the key idea(s), the evidence we have collected as part of the Building Evaluative Capability in Schooling Improvement (BECSI) Project, and advice that relates to these ideas and evidence. This advice is meant to help schools and clusters as they develop in their evaluative capability, and progress towards an optimal model as outlined in the first position paper (BECSI Position Paper 1: 'Towards An Optimal Model for Schooling Improvement'). As such there are various pathways that a school or cluster may take as they develop the knowledge and practice they need through cycles of inquiry and each section within the first position paper describes features of the likely developmental progressions.

There may be language used in this paper that is unfamiliar to some people involved in schooling improvement. To aid understanding of the points raised there is a glossary of statistical terms included at the end of the paper.

What do you need to do (and why)?

- **Test administration and types of tests**

In order to measure progress it is important to:

- Use nationally standardised tests
- Use a range of tests
- Make sure that each test is administered in an appropriate and consistent manner.

Why is it important to administer each test in a consistent manner? It gives us more confidence in the progress measurements calculated as the testing conditions, as well as the tests themselves, are as similar as possible.

The use of nationally standardised tests allows us to compare the progress we measure with an idea of what it is expected and therefore we can judge whether any progress that is being made is real, or large enough to have made a difference to the level of student achievement in the school or cluster.

Why do we need a range of tests? A useful analogy here is to think of measuring student fitness by the time taken to complete a task. Suppose that some schools decided to measure the time their students took to swim 100 metres, some to measure the time to run 500 metres, and others, the time to cycle 5 kilometers. All would be valid measures of fitness, all could be used to measure improvement in fitness, but none could be compared directly. Some of the measures might be less reliable than others. Suppose that the swimming schools used a variety of kinds of pool: some were salt-water, some fresh-water, and some of the fresh-water pools were heated to a constant temperature, while others were unheated outdoor pools. Some may even have timed students swimming in open water (sea or lake). Cycling times in Wellington might prove more variable than those in Christchurch, because of Wellington's variety in wind-speed and terrain. If all the running was to be done around a sports field, those times would be the most reliable (in both the lay and statistical sense of the word). If we want to measure a student's overall fitness, then measuring their running and swimming or cycling speeds would give a more complete picture than any one speed measurement on its own.

Returning to measuring student achievement in literacy or numeracy, we cannot measure progress from one assessment to another, at a later time, using a different assessment tool from the first (the reliability of such measures of progress is too low because of the different foci of the two tools). Comparable progress measures can be calculated where students using Test A at both times can be compared with students using Test B at both times. For example, we cannot combine a student's Supplementary Tests of Achievement in Reading (STAR) score at the beginning of the year with their Assessment Tools for Teaching & Learning (asTTle) reading score at the end of a year (this would be roughly equivalent to measuring their fitness by measuring swimming times at the beginning of the year and running times at the end of the year), but data from schools using STAR to measure literacy at two or more time points can be combined with data from schools using asTTle reading (or writing or mathematics), or any of the Progressive Achievement Tests (PAT) scores at two or more time points as an overall measure of progress (not in a specific learning area).

We use a range of tests to allow us to measure different things in the most appropriate way, for example, reading, spelling and writing all help to measure overall literacy, but require different tests. However we can compare results on different tests in the same content area in order to triangulate (or check) our findings (for how to compare different measures, see section on effect sizes). If two tests of a particular content area, for example reading, gave very different measures of progress this would alert us to think about what we're measuring, if we're using appropriate tools, and the strengths and weaknesses, as tested, of a particular student (or group of students). Conversely, if two tests of a particular content area gave very similar measures of progress we could be more confident in our measures.

However, in order to measure progress we must use the same test at two different times, for example, start and end of year, otherwise we cannot judge whether progress has actually been made. The choice of test should also relate to the valued outcome(s) as agreed and understood in the theory for improvement. (BECSI Position Paper 2: 'Theories for Improvement and Sustainability'). This applies not only to content area, but also to other choices, such as the age of students involved in the schooling improvement intervention.

We have found that some schools have changed the assessment tool they were using between time points either because they saw another assessment tool offered advantages over the one they used initially, or because they changed the focus of their schooling improvement initiative. At least one cluster collected baseline data for one set of year levels, but then decided to focus their schooling improvement on other years. In both these examples, measuring progress from the baseline was made impossible (although the changes may have had advantages for the schools that outweighed this disadvantage).

When we collected data for the value-added analysis of student achievement data strand of BECSI we received only STAR and asTTle data. From the Inventory phase we know that schools collect a wider range of test data than they use for the cluster work (Timperley, Parr, Hohepa, Le Fevre, Lai, dingle & Schagen, 2008); for example, both PAT tests and BURT Word reading test were mentioned in the Inventory Phase, as well as STAR and asTTle.

Good practice:

As stated before:

- Use nationally standardised tests to measure progress
- Administer each test in a consistent way
- Use a range of tests to help check progress findings

And,

- Use tests that measure the valued outcome(s) as set out in your theory for improvement
- In order to gather a richer picture of the impact of your schooling improvement intervention, both schools and clusters that are using a wide variety of nationally standardised tests could compare the evidence from these tests

Data collection

It is important that all student data are collected and stored accurately. In order to measure progress the test score needs to be stored alongside important demographic variables such as gender and ethnicity. These variables are best extracted or exported from the School Management System (SMS), as they are then most likely to be correct, consistent and up-to-date. It is important that schools check the accuracy of their SMS to aid this process. As a school will be using the data for both formative planning and classroom use, as well as summative progress and impact measures, we expect schools to store subtest information as well as total, scale and stanine scores.

We found that student achievement data were collected by schools and often collated at a cluster level. The processes used for collecting and entering these data and for checking the accuracy of these data are improving in schools for single assessments (at a single time-point). Previously we had found that there were many mistakes being made in recording correct scores and conversions (into stanines for example), and much data checking and cleaning was needed before analysis could be begun. The data collected for the value-added analysis strand of work had far fewer typographical, manual or calculation errors.

Good practice:

- Keep checking accuracy of information
- Extract or export from the SMS where possible
    - Examples include: student ID number, last name, first name, date of birth, gender, ethnicity, ESOL identification, learning support identification, year level, and class identifier
- To this add a date of test (making it easy to double-check that the correct file is being used in analyses) and test results, recorded at subtest level
- Consider including more general information, particularly for use at the school level: teacher name and other information like number of years of experience, or at the particular school, or PD received; number of years the student has been at the school; flag if the student has left the school and returned (particularly if this happened more than once), etc.

Data storing for measuring progress

The only way that progress can be measured and therefore the impact of schooling improvement assessed is by storing test data to allow students to be matched over time in a longitudinal analysis. Obviously, student data can be stored for one year and compared at the start and end of that year, but in order to see progress over time it is useful to store student data for longer periods and track students across year levels. This is true at both school and cluster levels. When tracking students in a cluster of schools, it should be possible to track a student from a primary school through their intermediate schooling and into a secondary school, if a cluster of schools includes all these levels of schooling.

Most SMS in schools do not easily store data for longitudinal analysis so it may be necessary to create a database to do this. This can be done in a variety of ways, depending on the resources and skills available in each school or cluster. If a school regularly uses MS Excel, or similar spreadsheet package, or MS Access, or similar database package, then it is sensible to use one of these to store the data[2].

There are some “do’s and don’t’s” around collecting and storing data for longitudinal analyses. Tracking students across years can be difficult if variations on the students’ names are used at different times. Although some schools use an internal school ID number for their students these are not necessarily helpful, as it would seem that these can change with changing SMS system, or improvements in process.

The use of correctly recorded National Student Numbers (NSN) would make matching data within and between schools much easier.

The following list highlights useful ways to make storing data, and therefore being able to track students, easier for all involved.

Good practice:

- Student names should always be recorded in full: complete names, and all first names
- Real care needs to be taken with hyphens, spaces, and apostrophes, for example in Pasifika names
- Dates of birth are useful
- The Ministry of Education conventions for recording ethnicity should be followed
- Use of the NSN will make this task considerably easier (although great care needs to be taken when entering it), and names and dates of birth will still be helpful as backup.

What do you need to know (and why)?

- **Expected progress**

Measuring progress can be as simple as calculating the difference between two test scores assessed at different times. However, in order to calculate actual progress we need to take into account how much we expect that student to have progressed as they matured. All published nationally standardised tests in New Zealand have expected progress. This is the amount of gain we expect each student to make as they mature from one time point to the next. This way, if we know that between Years 4 and 6 we expect a student to make a gain of 100 marks on a test just by being two years older, and a particular student actually made 150 marks progress then the actual or real progress made by that student is 50 marks.

There are two types of expected progress typically found, one is the actual number of marks a student is expected to gain on average between two time points (e.g. asTTle and PAT); the second is the use of a standardised measure, for example stanines (e.g. STAR), where expected progress is to remain in the same standardised band or grouping.

asTTle has tests for Years 4–10. There are separate tests for writing, reading and mathematics, each measured on its own scale. These three scales cannot be compared directly. As each test is already measured on a single scale comparisons can be made between students in different year levels, although each year has a different amount of expected progress. For example, the mean score, nationally, in Year 5 for asTTle reading is 462 and the mean score on the same test in Year 6 is 489. Expected progress is then a gain of 27 on the asTTle reading scale (aRs). Therefore, a student gaining more than 27 from Year 5 to Year 6 has progressed more than expected.

The PATs are similar to asTTle, in that they range from Years 4–10, each year has a different amount of expected progress, and each test has its own scale and so tests cannot be compared directly. For example, the mean score, nationally, in Year 5 for PAT mathematics is 40.3 and the mean score on the same test in Year 6 is 46.4. Expected progress is then a gain of 6.1 on the PAT mathematics scale (patm). Therefore, a student gaining more than 6.1 from Year 5 to Year 6 has progressed more than expected.

STAR has tests for Years 3–9. There are separate tests for Year 3, Years 4–6 and Years 7–9. The possible maximum score on each test version increases from 45 (Year 3) to 50 at Years 4–6 and then up to 80 for Years 7–9. These “jumps” in possible total raw score means that raw scores from different year levels cannot be combined meaningfully . However, comparisons between students in different year levels made using stanines are meaningful as all stanine scores have a mean of 5 and standard deviation of 2. If a student is maturing at the expected rate then they will score more marks (raw score) at each year level, but these gains will

[2] New generations of SMS software are planned to provide this capability.

translate into the same stanine at each year level. For example, a student scoring 23 at the end of Year 4 and 28 at the end of Year 5 has matured as expected and would be in stanine 4 at both time points. This means that a student who makes any stanine gain, for example from stanine 4 at the end of Year 5 to stanine 5 at the end of Year 6, has progressed more than expected.

- **Accelerated progress**

Schooling improvement is not just about making progress, but is about making enough progress to ‘close the gap’ between the students who perform at a nationally expected level and those that do not and therefore intervention is felt necessary. We call this closing of the gap accelerated progress. How much acceleration is needed varies across schools and clusters. However, a good starting point is to set targets that aim to move the school or cluster distribution from below average to average. This means that on average students in each school or cluster will achieve at national norms, but there will always be some individual students who achieve below this and some who achieve above this. Targets may be set to move the mean of the students achievement from below average to average, but it is important to consider the whole distribution as often you find that you need to move the lowest achieving students at a more accelerated rate than the students towards the centre of the distribution. Whilst doing this don’t forget the highest achieving students; they need to be achieving at the same level as before, if not doing even better.

- **Group analysis**

We often want to analyse different groups within our student achievement data. For example, are our boys doing as well as our girls? Is there one ethnic group that doesn’t achieve as well as the others? If we look at our lowest achieving students at the beginning of the year, say the lowest 20 percent, do they make as much progress as the remaining 80 percent? These are often very useful analyses to do as they can help with planning and assessing impact, especially in schooling improvement.

When comparing groups it is important to look at averages and their variance, that is the amount of spread there is for each distribution. There are different ways to do this, including interpreting a mean and its standard deviation, and looking at bar charts of distributions. We have to be careful that the groups we are comparing are large enough to be sure any patterns and trends we see are not just because of one good teacher, or an atypical cohort of students, or some other chance circumstance. A general ‘rule of thumb’ is that if each group has 100 students or more then we can be pretty certain that the differences we see are real (not just chance). Any groups smaller than 10 students are definitely too small to say anything meaningful, and fewer than 50 students may give misleading results (but it’s impossible to know when this happens). If your school or cluster is quite small, don’t be surprised of there is a lot of variation in your student achievement from year to year.

There is one other issue to consider when examining differences between groups. This occurs when we split the cohort of students into groups based on achievement, for example, we compare the lowest 20 percent of the cohort with the rest to see of they progress at different rates. The second and any subsequent assessments of these students will display something we call ‘regression to the mean’. Regression to the mean occurs whenever an extreme subset of individuals is selected (for example, the lowest or highest 20 percent based on their test scores at the first time of testing), and then measured again on a second occasion. It is certain that if the lowest-achieving subset is selected, some of the individuals will have been included in the group because they achieved lower-than-average scores but this was not a true reflection of their normal achievement level (they were having a ”bad day” if you like), and in the next test their scores are likely to be higher, closer to their “true ability”.

The result is that the mean scores of the group on a second occasion will be higher, irrespective of anything else that happens between the tests. The same is true for the highest-performing subset: in a second test the mean for the group will be lower, as some individuals (those included in the group by chance high scores, they were having a ‘good day’) are likely to have lower scores than in the first test. There are sophisticated calculations that can estimate the effect of regression to the mean based on how reliable the assessment tool is amongst other things, however, making such calculations are not necessary most of the time. The salient point to take from this is that regression to the mean occurs, and any differences you see between achievement groups should be treated with enough caution to allow that some of that difference is due to this phenomenon.

- **Effect size measures**

The most commonly used definition of “effect size” is an index that measures the strength of the association between one variable and another, for example between student achievement and an intervention. These indices take different forms depending on the measure being used. The most commonly reported effect size measure is Cohens *d* which is used to compare the means of two groups. Exactly what to use to calculate *d* requires careful thought, as there are different options, some of which have more advantages than others. A more detailed explanation of the calculation and interpretation of effect sizes, along with tables to help people avoid ‘traps’ and make the calculations easier, can be found in the ‘How much difference does it make?’ brochure written by Ian Schagen and Edith Hodgen, 2009 and available from either the NZCER or Ministry of Education websites (www.nzcer.org.nz or www.educationcounts.govt.nz/publications/schooling). Note that any effect size should be calculated along with a confidence interval to take into account the anticipated variability of the measures.

Once calculated, effect sizes and their confidence intervals can be used to compare progress measured using different tests. This will allow the triangulation to take place that will confirm or question the impact of your schooling improvement intervention.

Capability/Inter-dependence

Some of the ideas and suggestions outlined in this paper may challenge the skills of teachers and school leaders working in schooling improvement. As discussed in the first position paper (see previous reference) should the more sophisticated use of assessment data be left solely to school staff? There are three scenarios to consider here:

1. School staff are expected to assess students, store the data, and analyse the impact of schooling improvement with no external help. There are most certainly advantages for schools to be able to carry out all this for themselves. In the standards environment that schools are entering now there will certainly be a requirement for school staff to understand such measures, if not calculate them. The obvious disadvantages occur when sophisticated analysis is required that statistical professionals have spent time learning and training to be able to perform. Should school staff, who may not have any

statistical training, be expected to perform these analyses as well as carrying out the professional duties they have trained for?

2. The second scenario allows that some of the aggregation and more complex analysis are performed at the cluster level and not within schools alone. Again, as clusters in schooling improvement should be assessing the impact of their actions, there are advantages to this scenario. However, the disadvantage remains the same as the first scenario. Clusters are composed of school staff who may not have the required expertise or resources, and the professional developers and Ministry staff working with each cluster may have other skills that do not fit with the expertise required.

3. The third approach, possibly the one with the most advantages, is one of inter-dependence with external experts. Schools carry out the analysis that is most suitable for their needs, whilst facilitating the data storage required for aggregation to the cluster level. The clusters carry out analyses that are appropriate to assess the impact of schooling improvement, but now they draw on external expertise to guide and help them. This means that evaluative capability is built upon for all involved, and the expertise required is shared across all schooling improvement work to ensure that the best information is gathered and used from the assessment of students (in terms of measuring progress).

Other factors to consider

- **Standards framework**

The evidence provided by the project has shown that there has been improvement in how schools are collecting assessment data at a single time point, however, there are clear indications that storing these data for longitudinal use is not carried out as effectively. For any value-added work there needs to be a measure of progress, measured on individual students. If useable datasets of student achievement data are not produced and maintained then it will become impossible to measure and report on any standards work as there will be no baseline data for comparison. In any standards framework there is a need to show that a difference is being made to student learning and schools will struggle to do so without being able to measure the progress each student has made. For schooling improvement work based around school clusters this measurement also needs to be made at a cluster level.

- **Māori Medium Education (MME)**

An optimal model for schooling improvement in the context of New Zealand is one that can enhance outcomes in MME. Effective schooling improvement ensures that “achievement” is inclusive of the kaupapa of MME and kura kaupapa Māori. A model that is inclusive makes genuine space for kaupapa and aspirations.

One of the difficulties identified in Māori medium settings that participate in schooling improvement is the limited availability of appropriate assessments that capture the kaupapa of MME and kura kaupapa Māori needed in order to ensure a strong evidence base. However, where appropriate assessment is available, the advice given in the rest of this paper is pertinent with respect to measurement of progress. There are complexities in assessing bilingual achievement, as discussed in papers such as Rau (2005) especially when considering the language background of the students. It is important to record information about students’ language learning history because of these issues, for example, how many years has the student been educated in Māori immersion, has this been a continuous or broken amount of time, what level of immersion has the student experienced, what was the early childhood educational experience of the student. All these factors can then be considered when analysing student achievement data.

How this links to teachers’ formative practice

The analysis of student achievement with the view to measure progress does not replace the formative practice of teachers. Formative practice is based on inquiry, and schooling improvement is based on inquiry centred on student achievement data. As such, it is part of the same cycle as formative practice. The day-to-day use of student assessment in the classroom, for planning, and other uses remains important. This paper outlines how to measure, store data and perform analyses for determining student progress, especially in the context of assessing the impact of schooling improvement in schools and clusters. Some of the ideas conveyed here may also be helpful in analysing data for other uses, and assessment for progress should complement the formative analysis already being carried out in schools today.

Glossary of statistical terms

Average

A term loosely used to indicate a typical measure. More exact terms for an average are:

- **Mean**

  The total of all the measures divided by the number of measures. For example, the mean for a class test is the total of the test marks divided by the number of students writing the test.
- **Median**

  If all the measures are arranged in ascending (or descending) order, the median is the one right in the middle. For example, the student with the median mark has half the class getting a higher mark than they did, and the other half of the class getting a lower mark.
- **Mode**

  The measure that is most common. For example, for a STAR test, the mode nationally is stanine 5 as more students achieve a stanine 5 score than any other (this would not be true in an underachieving school, where the mode might be stanine 2 or even 1, if most students in that school achieved at that level).

Standard deviation

A measure of the typical variability in a set of data. It is a measure of how much the data points vary about their mean.

Confidence interval

When we estimate something like a mean test score for a cluster, or an effect size, it is helpful to have an indication of how accurate the estimate is likely to be. A confidence interval is defined by its upper and lower limits, and we can say with a known level of confidence (in the statistical sense) how likely it is that the true value of what we are estimating lies between these limits. For example, if an effect size of 0.30 has a 95% confidence interval of 0.15 to 0.45, then most probably the effect size really is within that range (it is unlikely to be either bigger or smaller).

Distribution

This refers to how measures are spread across their possible range.
For example, for STAR, the possible range is stanines 1–9, and the distribution is shown by plotting a bar graph showing how many (or what percentage) of students achieve in each stanine group. Distributions are often described as being *symmetric*, if as many students do well as do badly (the national norms are symmetric), or *skew*,

if more students do well (or badly) than otherwise.

Bar chart

Bar charts are useful to show a distribution across a few categores, for example gender, or ethnicity, or stanine scores. The heights of the bars show how many individuals fell into each category.

Effect size

A standardised measure (free of the original units of measure like temperature, or asTTle score) of the difference between two groups that takes into account the variability in the measures, but not the size of the groups. It can be used to mak e comparisons with other similar measures.

Expectation

In this document the term is used to refer to what student progress we would expect within a particular length of time, based on the national norms for a test.

Maturation

In this document the term is used to refer to the expected growth in student achievement we would expect within a particular length of time.

Progress

A change in the ability of students over time.

Accelerated progress

Greater progress than expectation taking only maturation into account, or when students show greater improvement their achievement levels within a particular length than we would otherwise expect.

Regression to the mean

The tendency of measures, over a long period, to become more average. A tall father is more likely to have sons who are shorter than him than not. A student performing exceptionally well in one test is more likely to have slightly lower marks in the next. It becomes an issue when a group is selected *on the basis of their extreme measures* (the best/worst 20%, say), because that same group is almost certain to have more average measures on the next occasion just due to regression to the mean. It is therefore more difficult in the case of these subgroups to tell whether any progress observed is due to an intervention, or solely regression to the mean.

References

Timperley, H., Parr, J., Hohepa, M., Le Fevre, D., Lai, M.K., Dingle, R., & Schagen, S. (2008). Findings from the Inventory phase of the Building Evaluative Capability in Schooling Improvement Project. Report to Ministry of Education.

Rau,C. (2005). Literacy Acquisition, Assessment and Achievement of Year Two Students in Total Immersion in Maori Programmes. The International Journal of Bilingual Education and Bilingualism. Vol. 8. No. 5 pp 404-429