It Worked!

Supporting robust data analysis in a CoL

DARCY FAWCETT

This Assessment News article introduces readers to a statistical approach to making sense of student assessment data in order to help teachers understand whether or not changes in practice have made a difference to learning. It Worked! is the brainchild of Darcy Fawcett, HoD Science at Gisborne Boys' High School, and Across-School Teacher for the Turanganui-ā-Kiwa Gisborne Kāhui Ako Community of Learning. In this article, Darcy explains a range of ways that schools can apply research-based methods to generate data stories which can illustrate evidence of learning. Valid assessment data stories allow teachers to more effectively evaluate, improve, and share practice. The article concludes with insights into how data stories can support teacher inquiry, including the expertise, procedures, and collaborative relationships which can help.

Introduction

In my role as head of department (HoD) of science at Gisborne Boys' High School, I wanted to better understand student achievement and whether changes we had made in our teaching and learning were actually leading to improvements in student outcomes. To do this in a way that was statistically sound and valid, I developed a process for analysing data in schools, which I have called 'It Worked!' In 2018 I was awarded one of The Education Hub's inaugural Bright Spots Awards for It Worked! I have been able to implement It Worked! in all the departments at Gisborne Boys' High School. Together with my position as an across-community teacher in the Turanganui-ā-Kiwa Gisborne Kāhui Ako Community of Learning (CoL), I am also rolling the process out into primary, intermediate, and secondary schools across our CoL.

This article describes the research-level methods of data analysis being utilised to analyse school data. I illustrate how I work with teachers and leaders to support them in understanding the analyses and then how to use these to continually improve their practice. Although It Worked! is still in its early days, there is good reason to believe that it is having a substantial impact on how teachers and schools are utilising data and evidence to inform their work.

https://doi.org/10.18296/set.0144

An overview of It Worked!

How can you evaluate the effect of a teaching initiative? What data do you need, and how should you analyse them?

The It Worked! approach compares the assessment outcomes of two groups of students using assessment outcomes such as curriculum levels, PAT scores, e-asTTle levels, NCEA credits, grades, and endorsements. I use research-level procedures to extract meaning from these types of achievement data. The methodologies and analyses of It Worked! have been imported from educational research. They give our inquiries and conclusions greater statistical validity (truthfulness, reliability, usefulness, and the like). If research-level evidence shows that student outcomes have been enhanced, then the initiative worked (hence It Worked!). If not, the initiative did not work. These findings can help individual teachers or departments understand if an initiative worked, but their validity also means we can confidently share our findings across our community to the benefit of all.

Research-level methodologies

To evaluate the effect of a teaching initiative we need to compare two cohorts of students: one group who received the new teaching/learning and one who did not. To do this, we use either the experimental method or the quasi-experimental method. Similar to the natural sciences, in the experimental method we create two equivalent groups of students, teach the control group as normal, and teach the experimental group using the new teaching/learning activity. Although this gives a direct evaluation, the experimental method is problematic in schools because there is a range of variables in play. In the quasi-experimental method we use the new teaching/learning activity with our latest cohort of students and compare their results to previous years. This longitudinal comparison is much easier to arrange, but it assumes that the current cohort is a representative sample of the historical population.

If the students who experienced the new teaching/learning activity produced significantly better results than those who did not, then the new teaching/learning can be said to have enhanced that outcome. But we cannot make the claim that a change is "significantly better" on the basis of a raw comparison of averages or percentages. There is a range of performance each year and some years are better than others. Unfortunately, there is no hard and fast rule: "how much" is dependent on the characteristics of the two distributions in question. Luckily there are well-established statistical procedures that analyse the data and determine whether an improvement is significantly better, and, if so, how much better. On the down side, there is a different procedure for every type of comparison and dataset.

If research-level statistics seems daunting, don't worry. You don't have to understand the detail! Only one teacher in the community needs to be a statistician. Teacher-statistician is my ACT role in the Gisborne CoL. Thanks to computers, all data analysis can be automated and, with a little training, a single teacher-statistician can analyse and interpret all of their CoL's data. Everyone else just needs to learn how to interpret graphs and to include statistics such as the *p*-value in their data stories.

Telling robust data stories

A data story is the end product of my collaboration with teachers. It is a summary of our inquiry and conclusions. It drives our next steps. Of course, teachers are already telling informal data stories like "I'm going to use activity x instead of y because last time I used this activity the students achieved good grades." Although informal data stories provide a rich record of personal experience, their claims can lack validity. To increase the validity of our data stories, teachers need to use research-level methodologies and analysis.

In the It Worked! project, we are increasing the validity of our data stories by using the appropriate methodologies and analysis. We illustrate our findings using graphs, and support our claims with summary statistics such as averages, measures of spread, and the results of the appropriate statistical tests. Averages describe the middle of a group of results and measures of spread how the results are distributed. Remember, comparing raw averages is not good enough. Statistical tests provide us with powerful ways of describing the difference between the distributions of various cohorts. The tests for significant differences are most important and their meaning is summarised by the *p*-value. In teacher speak, the *p*-value gives the probability that any differences we find between the cohorts are the result of random variation (e.g., some years are more able than others). When the *p*-value falls below a critical value (e.g., 10%, 5%, or 1%), we can dismiss those random effects and claim that there are notable, significant, or highly significant differences between the outcomes of our initiative cohort and the rest. For example, if there is only a one precent chance the improvement in the grades is random, you can be pretty confident that your initiative has worked!

At first we focused on finding the *p*-value in NCEA analysis, but my primary colleagues are demanding more. The *p*-value only tells you that there is a significant difference between the cohorts. It doesn't tell you how big the difference is (e.g., how much more learning took place). If you want to describe the size of the difference—for example, to compare initiatives—you need to test the strength of association. These tests tell you the magnitude of the relationship between two variables. In teacher speak, the strength of association gives the size of the effect your initiative had on the assessment results. As the association increases between 0 (no association) and 1 (complete association), we can claim our initiatives have greater and greater effect on student outcomes.

The following data stories illustrate the power of these processes. All of them were generated within my CoL.

Story 1. Writing progress in a split-level primary class

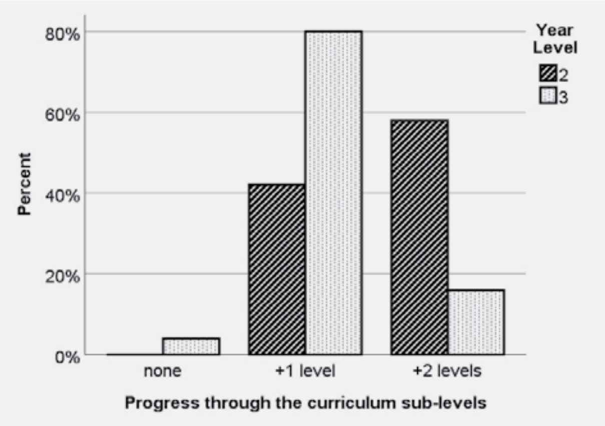

FIGURE 1. PROGRESS THROUGH THE CURRICULUM SUBLEVELS

A primary colleague in my CoL had a gut feeling that the Year 2 students in her split-level class were making more progress in writing than the Year 3 students. We used her overall teacher judgements (OTJs) for writing in 2017 and 2018 to find out. She recorded her OTJs using curriculum levels with two sublevels (e.g., Early Level 1, Level 1, Early Level 2, Level 2, etc.). The curriculum levels and sublevels form an ordinal variable (i.e., categories that have an intrinsic order). We had curriculum sublevel data for both February and November for 19 Year 2 students and 22 Year 3 students. We conceptualised the year's learning as progress through the curriculum sublevels (i.e., the difference between the end and start of the year). Progress then also forms an ordinal variable (no progress, +1 curriculum sublevel, +2 curriculum sublevels and so forth). This measure captured learning because it didn't matter where a student started, it's the progress they make over the year. For example, a student who went from Early Level 1 to Early Level 2 has made just as much progress (two curriculum sublevels) as a student who progresses from Level 2 to Level 3.

The bar graph shows the pattern we found; the cross-hatched bars represent the progress of Year 2 students and the light grey bars that of Year 3 students. The horizontal axis shows the amount of progress made during the year. The height of the bars illustrates the percentage of students who achieved that amount of progress. When looking at bar graphs, the first thing you notice is the highest bar. This gives the most common outcome (which is the mode average). As you can see, the mode for Year 3 students is at +1 curriculum sublevel, whereas the mode for Year 2 students is at +2 curriculum sublevels. That is, most Year 2 students made two curriculum sublevels progress, whereas most Year 3 students only made one curriculum sublevels progress. In general, it looks like Year 2 students made more progress than Year 3 students.

To increase our statistical validity, we utilised the Linear-by-Linear Association statistical test. This is the appropriate test for comparing two groups using an ordinal measure such as our progress variable. This test showed the differences in progress made by Year 2 and Year 3 students were statistically significant ($p = 0.006$) (i.e., the chance of this pattern happening randomly was a tiny 0.6%). My colleague's gut feeling was correct! Her Year 2 students really are making significantly more writing progress than her Year 3 students. What would you do if this was your class? And how might you know if your initiative had been successful?

Story 2. Comparing reading initiatives in an intermediate school

A local intermediate school groups their students and teachers into 16 hubs of between 25 and 91 students. This enables collaboration amongst teachers and encourages innovative teaching/learning. In 2017 there was a school-wide goal on strengthening reading, but the teachers in each hub could choose their collaborative approach to achieving this goal. The teachers collected data from various reading assessments to produce an OTJ for Terms 1 and 4. They were using curriculum-level data with three sublevels (e.g., Beginning Level 2, Middle Level 2, At Level 2, and so forth). They wanted to know whether their students were making progress and, if so, did the progress vary with teaching approach.

Their curriculum sublevels also form another ordinal variable. However, I couldn't calculate progress in the same way as for the primary school writing story. The dataset they gave me did not link each student's Term 1 data to their Term 4 data. I just had two sets of curriculum levels for each hub. But there is more than one way to skin a cat. Instead of calculating progress, I compared the shape (distribution) of the Term 1 data to that of the Term 4 data. Learning will be represented by Term 4 distribution being pushed (skewed) towards the higher curriculum levels when compared to the Term 1 distribution. The bar graphs below show the distribution of reading levels for Hubs X (91 students) and Y (63 students) in Terms 1 and 4.

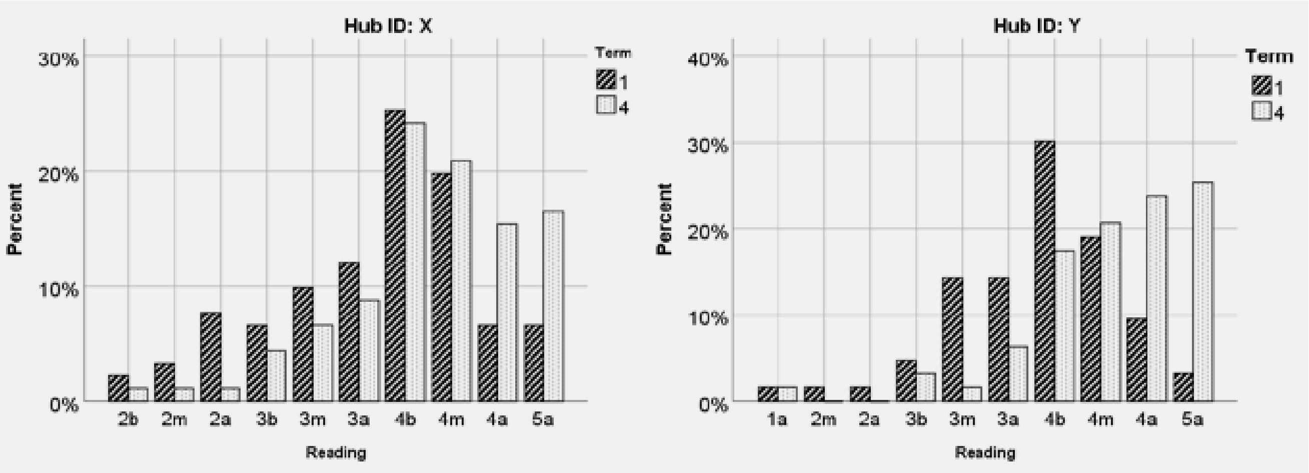

FIGURE 2. DISTRIBUTION OF READING LEVELS FOR HUBS X AND Y

Visually, it looks like learning has taken place in both hubs: the Term 4 light-grey bars show the expected skew towards the higher curriculum levels when compared to the Term 1 cross-hatched bars. The Linear-by-Linear Association test confirms that the Term 4 curriculum levels are significantly better than the Term 1 curriculum levels for both Hub X ($p = 0.001$) and Hub Y ($p < 0.001$). There is no more than a 0.1% chance of these improvements happening randomly.

It also looks like more learning took place in Hub Y as the highest bars (mode) for Hub Y are at the 5a end whereas for Hub X the mode is still in the middle above the 4b. To confirm our visual impression we need to use the Gamma measure of association. In this context, Gamma is used as a measure of progress through curriculum levels and describes the amount of skew on the graph ($g = 1$ would mean huge progress, $g = 0$ no progress at all). This test confirms our visual impression that the progress was larger in Hub Y ($g = 0.572$) than in Hub X ($g = 0.343$). Although I am encouraging both Hub X and Hub Y to share their reading approaches with the rest of the school, Hub Y has good reason to say their approach is better!

Story 3. Closing the gap in NCEA Level 1 science

In my school, Māori students used to earn fewer NCEA Level 1 science credits than Pākehā students. This is shown on the population pyramid for 2013.

The vertical axis shows the total credits earned and the horizontal axis shows the percentage of students earning those credits. The light-grey bars represent the credits earned by the 86 Māori students and the dark-grey bars those earned by 84 Pākehā students. The length of the bars shows us that the mode for Pākehā is at 20 credits whereas the mode for Māori is at 16 credits. The dark-grey bars are skewed upwards compared to the light-grey bars. Although there is a large range (the difference between the top and bottom credits) for each ethnic group, it certainly looked like Pākehā earned more credits in 2013 than Māori students did.

NCEA credits represent a different type of assessment data than those discussed so far. They are numbers on a scale and are non-parametric variables. (Parametric variables are numbers that are evenly distributed about the middle like the old School Certificate percentages.) The appropriate test for checking for significant differences between two cohorts when using non-parametric variables is the Mann–Whitney U Rank test. This test revealed that, in 2013, Māori students (median = 12) earned fewer credits than Pākehā students (median = 16) and that this difference was significant ($p = 0.002$). The median average describes the credits earned by the middle student when the credits earned are ranked in order. The only good news was that the R^2 test of association showed that the relationship between the credits earned and ethnicity is relatively weak ($r^2 = 0.057$). This tells us that the results for Māori and Pākehā students are not skewed in opposite directions.

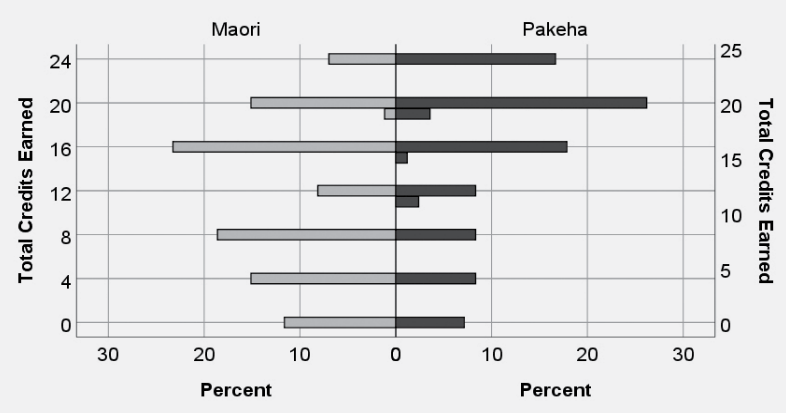

FIGURE 3. LEVEL 1 SCIENCE 2013: TOTAL CREDITS EARNED

Since 2013, we have continually refined our Level 1 science curriculum and used the total NCEA credits earned by Level 1 science students to evaluate the impact of our efforts. The line graph in Figure 4a shows how the mean average credits earned by Māori (light-grey dots) and Pākehā (dark-grey dots) has changed over time. The mean average is calculated by dividing the sum of the credits by the number of students. The size of the dots indicates the relative number of students (bigger is more students). Although there were disappointing results in the earlier years (look at the dip in 2014), we resisted deficit theorising. Our curriculum has become more responsive to the needs and aspirations of our students. As you can see by the converging lines, Māori and Pākehā now achieve on average the same number of credits.

The population pyramid graph (Figure 4b) shows the distributions of total credits earned in 2017. Visually, it looks like there are no big differences in the distribution of credits earned by Māori and Pākehā. The length of the bars at each credit level seems reasonably similar. The Mann-Whitney U Rank test showed that Māori and Pākehā averaged the same credits (median = 16) and any differences between the two groups are insignificant (*p* = 0.74). The test confirms that there is a 74% chance that any differences you might spot are random.

What we have learnt so far

A lot of work has gone into producing these stories and many others like them. Here is what we have learnt about using data stories to support teacher inquiry.

Asking meaningful questions is the first step. You can only answer questions for which you have the relevant data. Each of the above scenarios linked achievement data to some other type of variable (e.g., cohort, initiative, or ethnicity). Ask why a question is worth asking. The answer generally will boil down to:

- Checking a hunch—Did Year 2 students actually make more progress in their writing than the Year 3 students in the same class?
- Comparing initiatives—Was one reading approach more effective than another?
- Evaluating impact—Did the revised curriculum improve student outcomes? Do Māori still achieve fewer credits than Pākehā?

Being systematic about data collection is critical. In each of these cases, the question could only be answered because data about the relevant variables had been collected in the first place. But if you want to make comparisons, data have to be collected consistently. For example, if you want to describe year-on-year trends (a longitudinal analysis), the same data have to be collected in the same way each year. The further back your longitudinal data goes, the more certain you are to have included the full range of students in your community. This smooths out the good/bad years, provides greater sensitivity, and can generate convincing stories about the impact of changes. On the other hand, be careful to exclude radical external changes such as the NCEA realignment. Including these will mean your longitudinal comparisons are nonsense.

Every Kāhui Ako needs a data geek and time to share. Two things have allowed our CoL to investigate and tell data stories. I developed the necessary skills to analyse data through my own studies and passion for educational research. But I can only share my passion and support other teachers because I am an ACT. This gives me 10 hours a week to support colleagues across the CoL. And because I won a Bright Spots Award, I have money to release colleagues from their classes. Rather than after school, we now work together during the day to ask and answer meaningful questions about our students.

Linking datasets can be tricky. Ideally, we want to look at the effect of our teaching at both the individual level and at the cohort level. To do this, you need to link

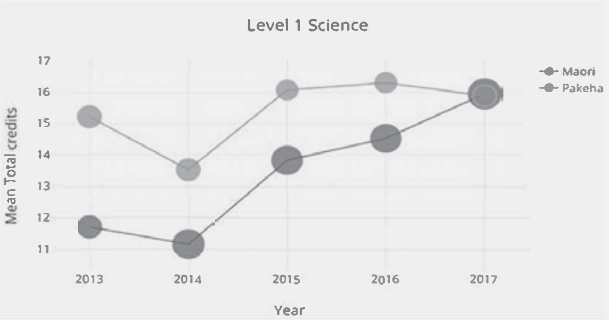

FIGURE 4A. MEAN AVERAGE CREDITS: CHANGE OVER TIME

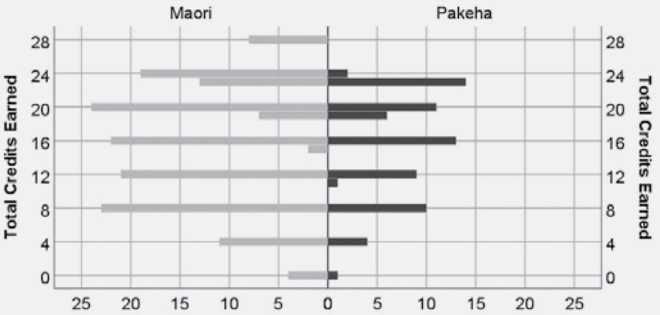

FIGURE 4B. TOTAL CREDITS EARNED BY MĀORI AND PĀKEHĀ

the individual student to his or her assessments across time. (Remember how I was able to calculate progress for my primary colleagues, but not for my intermediate colleagues.) As well as looking at progress across time and subject, we want to be able to disaggregate achievement data by demographics (ethnicity, gender, age, etc.). To do this, you'll need to be able to link all these variables to the various achievement datasets. The good news is that our student management systems already store all these types of data. The bad news is that extracting linked datasets is currently very difficult, if not impossible.

No-one should have to copy and paste multiple datasets to create a complete record. This is a major limiting factor when investigating student learning. It is time consuming to copy and paste datasets into the same record. There is also a real risk that errors will creep in. There is a better way. The funding and support I have received from the Bright Spots Award and the ongoing support from the amazing Robyn Caygill and Marian Loader in the Ministry of Education have allowed me to work on this challenge on behalf of the teaching collective. I have developed a new process for extracting a school's NCEA history split by subject. I cannot stress how important this is for making longitudinal data analysis viable. I am now working with NZCER to develop a new process for extracting a school's PAT and STAR histories. I've got my sights on e-asTTle. But the real challenge is going to be engaging with the student management system providers because we want each school dataset to talk with all others.

No-one should analyse data by hand. I started out analysing my department's NCEA data using drop-down menus. Now I write scripts (code) that automate the desired analyses. What used to take days, now takes a minute. This means I can share my skills and provide colleagues with the methodologies and analysis that ar necessary to give our inquiries and conclusions greater statistical validity. Senior secondary colleagues all want the same NCEA analyses—credits, grades, and endorsements, compared by cohort and disaggregated by gender and ethnicity. Junior secondary, intermediate, and primary colleagues are asking about curriculum levels, progression, and learning. All colleagues want to compare one cohort with another and to find out whether student outcomes have improved. When you get down to it, the underlying statistics and coding of these analyses are quite similar. Although conceptualising and scripting the desired statistics does take time, variations do not take too much time to code.

Make the findings count. Focus colleagues on interpretation and evaluation. Not everyone needs to be a statistician. I perform this function for my CoL. As well as scripting the statistical analysis, I sit alongside my colleagues and together we extract the key understandings from the statistical output. I understand the stats; they understand the context. Together we write the report. The stories above are summaries of such reports. Our processes are continually being refined and modified as a result of our collective experiences.

Where to next?

Our CoL is now discussing some wider issues that have arisen during our explorations. The following questions give you an idea about the sorts of issues that are opening up as a result of It Worked!

- What do "curriculum levels" mean in different contexts? How are levels determined and coded?
- Are we clear about the difference between causation and correlation? How do we ensure we don't over-claim on the basis of the impacts we find?
- Can It Worked! be scaled up outside our CoL? What structures might be needed to do that?
- How do we get leadership buy-in? How might individual classroom teachers engage with this work?
- What are the limitations of this work? What questions can't these types of analyses answer?
- How will my colleagues respond when it doesn't work? What structures and support can best support teacher collaboration and learning?

Acknowledgements

Thanks to Nina Hood at the Education Hub, Rose Hipkins at NZCER, and Robyn Caygill and Marian Loader at the Ministry of Education.

Darcy Fawcett is Head of Science at Gisborne Boys' High School and Across-School Teacher for the Turanganui-ā-Kiwa Gisborne Kāhui Ako Community of Learning. He was a Woolf Fisher Teaching Fellow 2016 and Bright Spot Award Winner 2018. Darcy is a passionate teacher-researcher who has been inquiring into his own and others' teaching practice for 26 years in New Zealand and the United Kingdom. Darcy has a BSc (Physics and Maths) and DipT from Waikato and an MA in Science Education with Distinction from King's College London.

Email: darcyf@gisboyshigh.net