The myth of New Zealand's declining educational performance

Prof. Gavin T L Brown, *The University of Auckland*

Much has been made of the gradual decline of New Zealand's rank order position on the OECD PISA tests. When PISA was first deployed New Zealand tended to score in the first 5 to 10 countries of the world. Now, however, New Zealand is ranked near the middle. This has been used to suggest that there is something wrong in our educational system. I wish to address some of the myth behind such claims.

Is New Zealand getting worse?

In the 1990s the Ministry of Education introduced an 8-level curriculum framework with the expectation that by the end of Year 8, students would have completed Level 4 and be ready for Level 5 at the start of high school[1]. So the question is simple: *Have we got worse since that framework was brought in?*

In 2006, the Ministry of Education produced a summary of student achievement in mathematics, reading comprehension, and writing[2]. They created an overview using information from the *Assessment Tools for Teaching and Learning* (asTTle) research and development team that I helped lead from 2000 to 2005, the *National Education Monitoring Project* (NEMP) run by Otago University, international test studies administered by the Ministry itself including (*Trends in International Mathematics and Science Study* (TIMSS), the *Progress in International Reading Literacy Study (PIRLS),* and *Programme for International Student Assessment (*PISA))**,** and from the *National Certificate of Educational Achievement (NCEA)* run by the New Zealand Qualifications Authority.

Those studies showed that performance levels in reading comprehension at years 5, 8, and 10 had not increased since earlier data in the 1990s. Likewise, performance in writing had not changed in years 4 and 8 between 1998 and 2002. Similarly, performance in mathematics showed small or no meaningfully significant changes between the middle 1990s and 2003.

The *National Monitoring Survey of Student Achievement* (NMSSA) run by the University of Otago and the New Zealand Council for Educational Research allow us to determine if the last 20 years has seen any improvement.

- The asTTle system found that 30% of Year 8 students performed in writing at Level 4 or above; this is the level expected by the curriculum for Year 8 students. More recently the NMSSA reported in 2019[3] that only 35% of Year 8 students scored at Level 4 or above.
- In 2019, NMSSA[3] reported 56% of Year 8 students were reading at Level 4 or above, while asTTle found just 25% of Year 8 students at the same level. This is a noteworthy increase.
- In mathematics, asTTle showed Year 8 students almost equally split between Level 3 and Level 4, while NMSSA[4] reported that only 45% of Year 8 students achieved at Level 4 or above, with most of the remaining students at Level 3.

Across these three key subjects, we have not gotten worse at all in terms of what we think is important for students to learn by the end of primary schooling. It is not valid to conclude that NZ children are achieving worse now than before[5].

But what about the NZQA literacy/numeracy tests?

It has recently come to the public's attention that the Ministry of Education (MoE) and the New Zealand Qualifications Authority (NZQA) disagree as to why so few students are passing the recently created compulsory tests in reading, writing, and mathematics. If we look at the NZQA test results and compare those to other sources, we can determine if there is an issue.

Evaluation Associates ran an evaluation of the new literacy and numeracy digital tests that are meant to be a co-requisite for earning NCEA[6]. The tests were run in Years 9-11 with 85% of participants in Year 10, the year before undertaking NCEA Level 1. The Ministry advised schools that the minimum level of readiness for students to undertake a Literacy and Numeracy assessment was late Level 4/ early Level 5 of the New Zealand Curriculum (NZC). Note that according to the curriculum framework that is the standard for Year 8 entering into Year 9, not the standard for Year 11 which is Curriculum Level 6.

Overall, the pilot reports that nearly 80% of students were at or above Level 4 Advanced on an e-asTTle test when they were tested, indicating they had met expectations for the end of primary school. This has strong similarity to the asTTle norms[2]:

- In mathematics, the vast majority of Year 10 students were in Level 4 or above, but Level 5 was not reached on average until Year 11.
- In reading, about two-thirds of Year 10 students were in Level 4 or lower. Level 5 was not reached on average until Year 11.
- In writing, about 35% of students in Year 10 were at the top of Level 4 or better. Even the Year 11 average was still in Level 4.

What this tells us that the sample used in the NZQA testing, was not much different in performance to the ability of students measured in 2002-20004 by asTTle. The 2022 study reported that two-thirds of students met the standard for reading and mathematics, but only half of students met the standard in writing. We might take this result in light of the e-asttle data to suggest that the standard for NCEA Level 1 is being set somewhere around the beginning of Level 5 rather than Level 6. Nonetheless, the results of this testing seem consistent with much earlier data. Does that mean NZQA made bad tests? I think not. Rather, the NZQA tests seem to be an important “canary in the mine” warning us that there is something wrong in our school system. While it could be unfamiliarity with formal digital testing, the warning may be about our teaching. Nonetheless, if we don’t like the results, we shouldn’t ignore the test data. Instead, we need to consider what has gone wrong in how we have, or haven’t, taught our students.

What about frequent testing and reporting?

We were introduced to the idea of frequent formal reporting when National Standards were introduced in 2010[7]. It is not a bad thing to tell families regularly where students are and use formal mechanisms to generate data. Parents want to know if their own children are struggling[8] and unfortunately, much of what teachers report to parents prevents them from knowing how well they are really doing[9]. Using standardised tests to monitor progress is a way around relying on teacher intuition and gut-feeling, a practice that was common in primary school teacher assessment[10]. But how often should tests be used?

Most schools are familiar with the tradition of the Progressive Achievement Tests (PATs). PATs were designed to be administered at the start of the year to get a sense of overall readiness and performance of a cohort. To minimise practice effects a parallel but different test was created for use in the following year. Hence, PATs were designed for use once-a-year.

In contrast, the e-asTTle system has banks of 1000s of questions in reading comprehension and mathematics. The software has an algorithm that is programmed to minimise re-use of previously administered items and all items have been calibrated to the same scale. This means that when a new test is created it is likely not to be identical to a previous test even if the same level and content switches are retained. This means that as teachers administer e-asTTle tests, there is a reduced

likelihood of an artificial 'practice' effect in having seen the same items before. This means reasonably frequent testing can take place. The original asTTle manual[11] recommended not more than four tests per year in any subject so that teachers spend time on teaching the new material students need in order to make progress, rather than conduct frequent testing.

So using e-asTTle tests up to 4 times per year makes sense, especially if teachers teach to the hard learning objectives that students struggled with between tests. Improved scores on e-asTTle come only when students start answering correctly harder questions, so repeated tests should always be slightly harder every time because students learn constantly. Teachers and leaders can have confidence that test item practice is minimised, so gains represent learning not memorisation.

Curriculum Level Width

With the 1994 curriculum framework, the first five levels of the curriculum were designed as 2-year progressions. The National party proposal is to report every year on progress saying the current move to three year bands of expectations doesn't give enough nuance to progression. In a sense that is correct even within the current system.

As such multi-year levels are not a bad thing because substantial and deep learning takes time. However, for children to remain at the same curriculum level for 2 or 3 years despite making progress within that level is discouraging for all involved. It is important and motivating to know that one's efforts are paying off in a positive trend even if the increments are small. How else do athletes know if they are getting stronger, faster, and higher but by recording and tracking their performances over time? So tracking and reporting progress is a good thing.

For those reasons, the e-asTTle system breaks each curriculum level into 3 sub-levels called Basic, Proficient, and Advanced. These terms give a positive spin on what students can do and where they are at within each zone of learning. The e-asTTle system also provides a numeric score system based on item response theory statistical algorithms that allow for even finer grained analysis of progress. A gain of 22 points in any subject is a statistically significant shift in either direction. A difference between students or times of less than 22 points simply reflects the random variation we can expect in human performance and the ability of tests to measure human learning.

So, we already have a broad range curriculum level framework in our school system, whether those be 2 or 3 years. We already have tools in the e-asTTle to monitor change at quite a narrow level. We can already provide for sufficient information to identify and track progress annually, without having to change the curriculum framework or be worried by those changes..

Why do we look so bad on International Large-Scale Assessments?

I want to consider some issues that relate our declining rank on ILSA measures, especially the age 15 Programme for International Student Assessment (PISA) tests. Here I suggest that our apparent decline is not something we should be concerned about, at least in terms of making policy concerning curriculum, teaching, or teacher education.

Sampling.

New Zealand, like many other official OECD members, conducts diligent sampling of the whole nation. That means remote rural, suburban, provincial, and high-density urban populations are included in the PISA sample in proportion to the population. While New Zealand is only a small population of just 5 million, it is very diverse in terms of socioeconomic status, educational backgrounds, linguistic and cultural features, and has a strong commitment to the indigenous

population. While New Zealand might not have been a colonial power, like many of the other European and North American countries we have large migrant and refugee populations that have legal right to be part of the school system and thus are included in the PISA test samples.

However, not all of the currently highly ranked jurisdictions on the PISA scale follow similar practices or have similar conditions. For example, Hong Kong, Singapore, and Macau are all in effect city states that do not allow easy comparison to nation-states[12]. It might make sense to compare such cities to Wellington or Auckland but not to a whole nation. Indeed, recent research that contrasted Shanghai's 2012 PISA results with New Zealand in the same year suggested that the gap between the two jurisdictions reduced considerably when only students in wealthy urban schools were considered[13].

This leads to a consideration of how China got to be so high on the PISA rankings. In the early days of 2009 and 12, Shanghai was the only part of China that was included and scored number 1 in the world[14]. At that time, China had a policy of providing health, education and social support services to citizens based on their *hukou* or household registration[15]. What that meant is that while citizens could move to big cities like Shanghai for work opportunities, they did not actually have legal rights to schooling and health services in Shanghai because their *hukou* was back at their village or town of origin. Consequently, none of the children without hukou rights could be included in the sampling frame for the PISA tests. This means that all the enthusiasm for finding out what Shanghai had done to reach number one in the world in 2009 and 12 was based on a sampling system that was legitimate in China but unimaginable in our own nation[16]. Let me reiterate, China did not cheat! They simply had a different set of rules about who was in the Shanghai school system.

More recently, China has expanded its coverage to include now the provinces of Beijing, Shanghai, Jiangsu, and Zhejiang. In the 2018 Pisa results this group of Chinese students again ranked number one[17]. it is commendable that the Chinese sample is broader, but unlike Western nations, this sample still does not include children in the poor rural areas of this vast nation. The Uighur population of Xinjiang, the poor regions of Southwest Yunnan, and the vast middle of Hunan and Henan are excluded. Very little can be learned from a jurisdiction that does not include a fair representative sample of its entire nation. Hence, any reference to China or Shanghai outperforming New Zealand must be seen as inconsequential. it's simply not a matter of comparing apples with apples.

Test, test, test

A key feature of many Asian societies is the prominent role of testing. Some 3000 years ago a formal testing regime was created in China to identify and select candidates for entry and advancement in the Chinese civil service[18]. Without success on these examinations, children of the poor were condemned to a life of hard labour whether that be in the field or in the city. This has led to widespread acceptance that performance on tests is essential to guarantee advancement in society and work and even the imputation of moral virtue[19]. Consequently, contemporary school systems from India and Pakistan to China, Japan, and Korea are characterized by frequent high consequence testing. Important decisions are made, based on test scores, for grade promotion, selection to elite classes, or entry to the next level of schooling. This has led, specifically in China, to the common expression:

> **考考考**，**老**师的法宝；
> 分分分，**学生的命根**

[exam, exam, exam, teacher's magic weapon;
grade, grade, grade, students' lifeblood]

The testing regime of China meant that prospective teachers we surveyed did NOT agree that excellent teachers used assessment formatively to improve teaching[20]. A survey of >1000 practicing teachers found that teaching for examinations was positively associated with using tests to improve the teaching of students[21]. Indeed, Chinese[10] and Hong Kong[22] teachers tend to believe that regular assessment and testing leads to personal moral development of youngsters.

Under these circumstances, we might expect students will treat test scores as a kind of currency; a chance to purchase social and personal advancement. Being successful on tests brought pride and positive emotions to Hong Kong university students, but being tested frequently was not something that they thought was healthy or beneficial[23]. Chinese students who believed assessment was designed to improve teaching and learning had less anger and shame, while being held accountable for performance increased those negative emotions[24].

We also ran an experiment in Shanghai and New Zealand in which we studied the motivation and effort students said they would exercise on three kinds of tests; that is a no consequence research test, a country consequence test like PISA, and a high-consequence personal test. While we found that Shanghai students[25], like New Zealand students[26], would make more effort for personal tests than country tests, the gap for Shanghai students was much smaller than New Zealand students. It is as if New Zealand students have permission not to try on PISA tests, whereas Shanghai students must care about any test.

In contrast to Asia, New Zealand does not introduce testing, other than for diagnostic or formative purposes, until the 11th year of schooling[27]. Furthermore, given our qualifications system, it is possible to acquire enough credits for the NCEA without taking externally marked, final examinations. Hence, when a sample of NZ students faces a formal test associated with the PISA system, it is highly likely such children will be unfamiliar with the practice of taking tests. This is quite unlike youngsters in many other parts of the world where frequent testing is normal.

Successful Life, but not tests

Something we have to consider is whether we want a society and school system defined by frequent, high-stakes tests. It may be that our lower than Asia performance on PISA lies in what it means to be successful in our jurisdiction. For kiwis, there are many ways to become successful; art, music, sports, manual labour, technology and so on all lead to qualifications, incomes, status, and place in society. This is not the classic priority in China where university qualifications are highly regarded, even if they might lead to unemployment[28]. We do not abhor those not working in air-conditioned offices with clean hands and white collars. Indeed, with the various tax benefits associated with self-employment, plumbers, builders, sparkies, and so on are doing better than many university-educated workers.

Another related feature of our rich conception of success is an open policy, which we share with other western nations, of retraining and further training even without schooling success. Learning a new trade after school failure or unemployment leads to

social and economic success without being bound up with social prestige of school and higher education success. In the end, we care more about children being all that they can be whatever that is than their doing well on frequent testing.

Unfair and pointless comparisons

I am not alone in critiquing the underlying technical challenges that potentially invalidate the use of PISA by the OECD. We examined the performance of 55 jurisdictions on the 2009 reading test (Shanghai was #1) and found using factor analytic invariance testing that almost all jurisdictions were not comparable with Australia, New Zealand, Canada, and the USA[29]. Based on language, culture, socio-economic development, and educational practices, we recommended that comparisons within groups of similar nations could be justified. For example, the Nordic region (Denmark, Sweden, Iceland, Norway, and Finland) could justifiably compare themselves given the similarity of their systems, societies, and resources. Similarly, it would make sense to compare the English-speaking, child-centric systems of Australia, Canada, USA, NZ, and UK. But comparisons to East Asia would be nonsensical and invidious.

What I find especially annoying with PISA is that the leadership know about the many technical problems associated with the test, OECD persists in ranking all jurisdictions regardless of dissimilarity. To their credit, OECD does provide bands of jurisdictions that have statistically significant differences to the global average. In 2018, New Zealand outperformed Australia and USA, was almost the same as the UK, and was less than Canada[30]; so we are in the middle of nations like ours? The most important question to answer is simple. What do we want for ourselves? Adopt a narrow definition of success as test-performance or take a broad view of success as being good at whatever we enjoy.

[1] Ministry of Education. (1993). *The New Zealand Curriculum Framework: Te Anga Marautanga o Aotearoa*. Learning Media.

[2] https://thehub.swa.govt.nz/assets/documents/42490_The-Big-Picture-Student-Outcome-Overview-2001-2005_0.pdf

[3] https://nmssa-production.s3.amazonaws.com/documents/2019_NMSSA_ENGLISH.pdf

[4] https://nmssa.otago.ac.nz/reports-and-resources/mathematics-and-statistics-reports/

[5] Brown, G. (2023). Not worse, but not better: The challenge of raising student achievement in Aotearoa [Opinion]. *Sunday Star-Times*. Retrieved April 24, 2023, from https://www.stuff.co.nz/opinion/300859143/not-worse-but-not-better-the-challenge-of-raising-student-achievement-in-aotearoa

[6] https://ncea-live-3-storagestack-53q-assetstorages3bucket-2o21xte0r81u.s3.amazonaws.com/s3fs-public/2023-03/NCEA%20Te%20Reo%20Matatini%20me%20te%20P%C4%81ngarau_Literacy%20Numeracy%20Pilot%20Evaluation-Report%20Two_March%202023.pdf?VersionId=HhHaDJzIseIkwvvLnzn89JqqkKlJhbRv

[7] New Zealand Ministry of Education. (2009). *National Standards: Information for schools*. Learning Media.

[8] Robinson, V., Timperley, H., Ward, L., Tuioto, L., Stevenson, V. T., & Mitchell, S. (2004). *Strengthening Education in Mangere and Otara Evaluation: Final Evaluation Report* [Commissioned report to the Ministry of Education]. https://www.educationcounts.govt.nz/__data/assets/pdf_file/0006/9285/semo.pdf

[9] Hattie, J., & Peddie, R. (2003). School reports: "Praising with faint damns". *set: research information for teachers*(3), 4-9. https://doi.org/10.18296/set.0710

[10] Hill, M. (2000). Dot, slash, cross: How assessment can drive teachers to ticking instead of teaching. *set: research information for teachers*(1), 21-25. https://doi.org/10.18296/set.0779

[11] Hattie, J. A. C., Brown, G. T. L., Keegan, P. J., MacKay, A. J., Irving, S. E., Cutforth, S., Campbell, A., Patel, P., Sussex, K., Sutherland, T., McCall, S., Mooyman, D., & Yu, J. (2004, December). *Assessment Tools for Teaching and Learning (asTTle) Manual* (Version 4, 2005). Wellington, NZ: University of Auckland/ Ministry of Education/ Learning Media. https://doi.org/10.17608/k6.auckland.14977503.v1

[12] Bray, M., Adamson, B., & Mason, M. (2007). Comparative education research: approaches and methods. In M. Bray, B. Adamson, & M. Mason (Eds.), *Comparative education research* (Vol. 19, pp. 363–379). Springer. https://doi.org/10.1007/978-1-4020-6189-9_16

[13] Zhao, A. (2021). *Sources of Sample Bias in PISA: Selectivity and Effort Differences between Shanghai and New Zealand* [unpublished Ph.D. thesis, The University of Auckland]. Auckland, NZ. https://researchspace.auckland.ac.nz/2292/61616

[14] https://www.science.org/doi/10.1126/science.330.6010.1461; https://www.theguardian.com/news/datablog/2013/dec/03/pisa-results-country-best-reading-maths-science

[15] Chen, Y., & Feng, S. (2013). Access to public schools and the education of migrant children in China. *China Economic Review*, *26*, 75–88. https://doi.org/10.1016/j.chieco.2013.04.007

[16] https://www.washingtonpost.com/news/answer-sheet/wp/2014/03/20/so-how-overblown-were-no-1-shanghais-pisa-results/; https://www.brookings.edu/articles/lessons-from-the-pisa-shanghai-controversy/

[17] https://www.oecd.org/pisa/publications/PISA2018_CN_QCI.pdf

[18] Brown, G. T. L. (2022). The past, present and future of educational assessment: A transdisciplinary perspective. *Frontiers in Education*, *7*. https://doi.org/10.3389/feduc.2022.1060633

[19] Brown, G. T. L., & Gao, L. (2015). Chinese teachers' conceptions of assessment for and of learning: Six competing and complementary purposes. *Cogent Education*, *2*(1). https://doi.org/10.1080/2331186X.2014.993836

[20] Chen, J., & Brown, G. T. L. (2013). High-stakes examination preparation that controls teaching: Chinese prospective teachers' conceptions of excellent teaching and assessment. *Journal of Education for Teaching*, *39*(5), 541-556. https://doi.org/10.1080/02607476.2013.836338

[21] Chen, J., & Brown, G. T. L. (2016). Tensions between knowledge transmission and student-focused teaching approaches to assessment purposes: helping students improve through transmission. *Teachers and Teaching*, *22*(3), 350-367. https://doi.org/10.1080/13540602.2015.1058592

[22] Brown, G. T. L., Kennedy, K. J., Fok, P. K., Chan, J. K. S., & Yu, W. M. (2009). Assessment for student improvement: understanding Hong Kong teachers' conceptions and practices of assessment. *Assessment in Education: Principles, Policy & Practice*, *16*(3), 347-363. https://doi.org/10.1080/09695940903319737

[23] Brown, G. T. L., & Wang, Z. (2013). Illustrating assessment: How Hong Kong university students conceive of the purposes of assessment. *Studies in Higher Education*, *38*(7), 1037-1057. https://doi.org/10.1080/03075079.2011.616955; Wang, Z., & Brown, G. T. L. (2014). Hong Kong tertiary students' conceptions of assessment of academic ability. *Higher Education Research & Development*, *33*(5), 1063-1077. https://doi.org/10.1080/07294360.2014.890565

[24] Chen, J., & Brown, G. T. L. (2018). Chinese secondary school students' conceptions of assessment and achievement emotions: endorsed purposes lead to positive and negative feelings. *Asia Pacific Journal of Education*, *38*(1), 91-109. https://doi.org/10.1080/02188791.2018.1423951

[25] Zhao, A., Brown, G. T. L., & Meissel, K. (2020). Manipulating the consequences of tests: how Shanghai teens react to different consequences. *Educational Research and Evaluation*, *26*(5-6), 221-251. https://doi.org/10.1080/13803611.2021.1963938

[26] Zhao, A., Brown, G. T. L., & Meissel, K. (2022). New Zealand students' test-taking motivation: an experimental study examining the effects of stakes. *Assessment in Education: Principles, Policy & Practice*, *29*(4), 397-421. https://doi.org/10.1080/0969594X.2022.2101043

[27] Crooks, T. J. (2010). Classroom assessment in policy context (New Zealand). In B. McGraw, P. Peterson, & E. L. Baker (Eds.), *The international encyclopedia of education* (3rd ed., pp. 443-448). Elsevier.

[28] https://www.shine.cn/news/in-focus/2207027475/

[29] Asil, M., & Brown, G. T. L. (2016). Comparing OECD PISA Reading in English to Other Languages: Identifying Potential Sources of Non-Invariance. *International Journal of Testing*, *16*(1), 71-93. https://doi.org/10.1080/15305058.2015.1064431

[30] https://factsmaps.com/wp-content/uploads/2019/12/pisa-2018.png