Learning from Assessment: NAPLAN and Indigenous Students

Gilah Leder
Monash University
<gilah.leder@monash.edu>

Helen Forgasz
Monash University
<helen.forgasz@monash.edu>

In this paper we report trends over time of performance of non-Indigenous and Indigenous students on the Numeracy component of the NAPLAN tests. Possible links between student performance on the NAPLAN Numeracy test and the four components - Reading, Writing, Spelling, and Grammar - of the NAPLAN Literacy test were also explored. While the performance of both groups of students at all grade levels have remained fairly consistent over time, there were differences in the aspects of literacy most strongly related to the numeracy performance of the two groups.

In many countries large scale national tests are used regularly to monitor student achievement (Postlethwaite & Kellaghan, 2009). Since its introduction in 2008, such monitoring is done in Australia through the National Assessment Program – Literacy and Numeracy [NAPLAN].

> NAPLAN tests identify whether all students have the literacy and numeracy skills that provide the critical foundation for their learning, and for their productive and rewarding participation in the community. Students are assessed using common national tests in Reading, Writing, Language Conventions (spelling, grammar and punctuation) and Numeracy.
>
> NAPLAN tests broadly reflect aspects of literacy and numeracy common to the curriculum in each state or territory. The types of test formats and questions are chosen so that they are familiar to teachers and students. (Australian Curriculum Assessment and Reporting Authority [ACARA], 2011)

As described in the National Reports, NAPLAN tests are equated, that is, "results from NAPLAN tests in different years can be reported on a common achievement scale" (ACARA, 2012 p. iv). Minor fluctuations in longitudinal test results are expected, and it is only when "there is a meaningful change in the results from one year to the next, or when there is a consistent trend over several years, that statements about improvement or decline in levels of achievement can be made confidently" (ACARA, 2012, p. iv).

The putative benefits of the annual administration of the NAPLAN tests are listed on the website of the Australian Curriculum Assessment and Reporting Authority [ACARA]. They mirror those commonly put forward in the wider literature: assessment consistency across different constituencies, increased accountability, and a general driver for improvement. At the same time, the limitations of the tests are clearly recognized. NAPLAN tests are timed, cover only selected components of the mathematics curriculum, and indicate only how well a student performs on the test on a given day. In brief, criticisms of the NAPLAN testing regime:

> … range from the reliability of the tests themselves to their impact on the well-being of children. This impact includes the effect on the nature and quality of the broader learning experiences of children which may result from changes in approaches to learning and teaching, as well as to the structure and nature of the curriculum. (Polesel, Dulfer, & Turnbull, 2012, p. 4)

2014. In J. Anderson, M. Cavanagh & A. Prescott (Eds.). Curriculum in focus: Research guided practice *(Proceedings of the 37th annual conference of the Mathematics Education Research Group of Australasia)* pp. 389–396. Sydney: MERGA.

A test's most important characteristic, according to Nichols and Berliner (2007), is its validity – a multi-dimensional construct. To assess the validity of a test comprehensively, they argued that four measures, the 4Cs, are required. These are content validity: whether the test measures what it is intended to measure; construct validity: whether the test actually measures the concept or attributes it is supposed to measure; criterion validity: whether the test predicts certain kinds of current or future achievement; and consequential validity: the consequences and decisions that are associated with test scores. There is consensus in the literature about the importance and relevance of the first three measures. Traditionally, "test validity has been broken into three or four distinct types — or, more specifically, into three types, one of which comprises two subtypes. These are content validity, predictive and concurrent criterion-related validity, and construct validity (Messick, 1991, p. 7). As discussed below, the fourth dimension on Nichols and Berliner's (2007) list is a more controversial measure of validity, perhaps because, as they noted, this last measure relies more on personal values than do the other three aspects.

Information on the ACARA website of the scope, aims, and underlying rationale for the NAPLAN tests references the first three aspects of validity either directly or indirectly, but the fourth of Nichols and Berliner's (2007) dimensions, consequential validity, is given less attention. As mentioned above, this dimension is also considered less consistently in the broader literature. Some adhere to Shepard's (1997) position: "I argue ... that consequences are a logical part of the evaluation of test use ... My contention ... is that examination of effects following from test use is essential in evaluating test validity" (p. 5). Others, like the equally influential Popham (1997), consider that "the assembly of evidence regarding test-use consequences can be accomplished without considering such evidence to be a facet of validity" (p. 13). Importantly, despite their disagreement about the concept of consequential validity, both agree that the "social consequences of test use should be addressed by test developers and test users" (Popham, 1997, p. 13).

Our Study: Aims and Context

The Numeracy NAPLAN data for Indigenous students, and possible outcomes spawned by these data, are examined in this paper. Our explorations were shaped and constrained by Nichols and Berliner's (2007) notion of the judgements made and conclusions drawn on the basis of test scores. Our investigation had two distinct components.

- The first involved inspection of data in the publicly available annual NAPLAN reports to examine longitudinal trends of non-Indigenous and Indigenous students on the Numeracy component of the NAPLAN tests.
- The second investigation drew on more detailed NAPLAN data, provided by ACARA[1]. This source comprised both the Numeracy and Literacy NAPLAN scores for students from 26 schools[2] - allowing us to explore for possible links between student performance on the NAPLAN Numeracy test and the four components - Reading, Writing, Spelling, and Grammar - of the NAPLAN Literacy test. Whether

[1] We gratefully acknowledge the support of ACARA in providing the Numeracy and Literacy data.

[2] The 26 schools participated in the *Make it Count* project which ran from 2009 to 2012 under the leadership of the Australian Association of Mathematics Teachers. Approximately 40 schools in the metropolitan and regional locations around the country participated in this project aimed at supporting best teaching of mathematics for Indigenous students.

or not any links found were consistent for the groups of Indigenous and non-Indigenous students was of particular interest.

To provide a functional context for our explorations, a brief overview of recent findings about the mathematics achievement of Indigenous students is presented first.

Indigenous Students

Reports that Indigenous students, on average, perform well below their non-Indigenous peers on traditional measures of achievement are prevalent. For example, Thomson et al. (2012) reported with respect to TIMSS [Trends in International Mathematics and Science Study] that "In 1995 and 2003 the score difference between non-Indigenous and Indigenous students [at Year 4] was 69 and 60 score points respectively" (p. 51). In 2007 the gap increased to 91 score points but decreased in 2011 to 64 score points, because "the average score of Indigenous students had increased significantly from 2007, while that of non-Indigenous students remained unchanged" (p. 51). At Year 8, the gap favouring non-Indigenous students has remained at around 70 points. Data from PISA [Programme for International Student Assessment], the international test administered to 15 year-old students, paint a similar picture of lower performance for Indigenous students. Data for PISA 2012 revealed that:

> Indigenous students achieved a mean mathematical literacy score of 417 points, which was significantly lower than the OECD average (494 score points) and non-Indigenous students (507 score points). The mean score difference of 90 points between Indigenous and non-Indigenous students equates to more than two-and-a-half years of schooling.... (Thomson, De Bortoli, & Buckley, 2013, p. 18)

A mean difference between the two groups equating to some two-and-a-half years of schooling was also found for reading literacy. As well, for both the mathematical and reading literacy tests Indigenous students were under-represented at the higher end of the proficiency scale and over-represented at the lower end of the scale (Thomson, De Bortoli, & Buckley, 2013).

Two considerations are missing from these broad summaries. Student achievement levels are directly related to geolocation. For each grade level, and for both Indigenous and non-Indigenous students, the further from metropolitan cities schooling takes place, the lower is the mean NAPLAN achievement score (Forgasz, Leder, & Halliday, 2013). It is also worth noting that in Australia "Indigenous background is derived from students' self-identification as being of Australian Aboriginal or Torres Strait Islander descent" (Thomson et al, 2012, p. xxvii).

Explorations of NAPLAN Data

Exploration 1: Tracing the Performance of Indigenous Students using NAPLAN Annual Reports: 2008-2013

Data for Indigenous and non-Indigenous student performance on the NAPLAN numeracy tests at each grade level (3, 5, 7, and 9) for the years 2008-2013 were extracted from the annual national reports found on the web at http://www.nap.edu.au/results-and-reports/national-reports.html. The results were examined cross-sectionally and longitudinally. Whether there were apparent trends in the data was explored.

In Figure 1, the NAPLAN numeracy performance data (mean scores) for students in grades 3, 5, 7, and 9 for the years 2008-2013 by Indigenous status are plotted. Blue – or dark lines – represent the data for Indigenous students, and non-Indigenous student data are in pink – lighter lines. Each line has also been labelled to facilitate interpretation.

Several clear patterns are apparent:

- Performance over time is fairly consistent for students at each grade level - both for Indigenous and non-Indigenous students. Slight variations can be seen but with no consistent trend for improvement or decline apparent.
- At each grade level, there is a large performance gap between non-Indigenous and Indigenous students.
- The performance of Indigenous students is about two years behind that of non-Indigenous students (consistent with the TIMSS findings cited above). For example, on average, Year 7 and Year 9 non-Indigenous students are outperforming Indigenous students in Year 9.

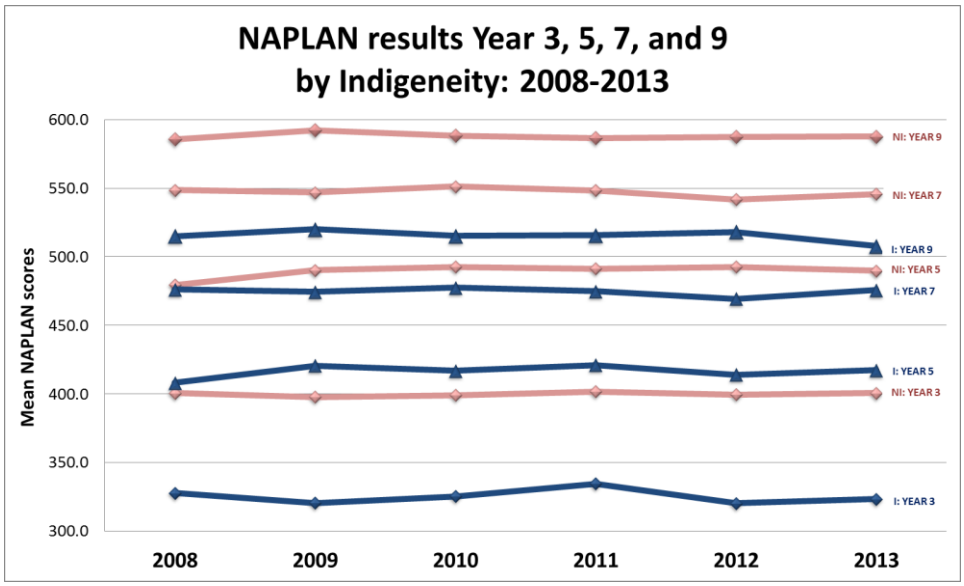

Figure 1. NAPLAN numeracy results by Indigeneity: Cross sectional (2008-2013)

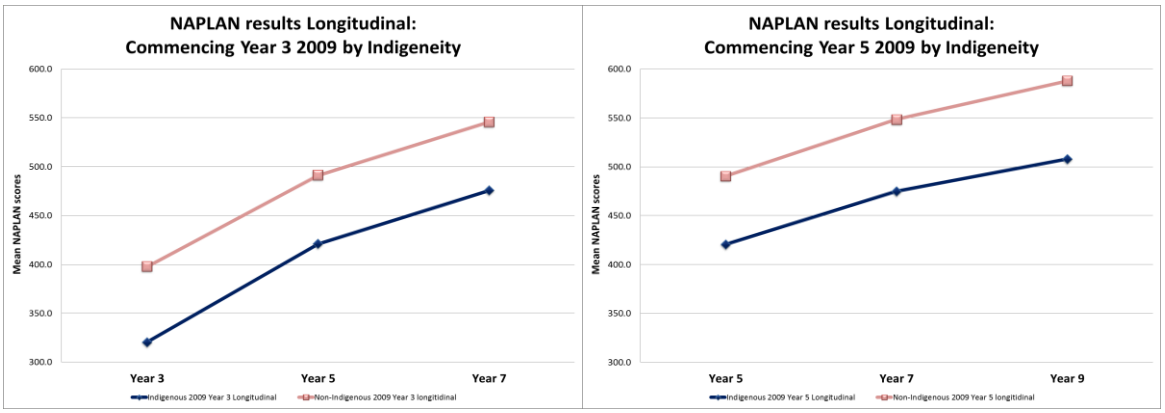

Figure 2. NAPLAN numeracy results by Indigeneity: Longitudinal (2009-2013)

With six years of NAPLAN data, it was possible to explore cohort data by Indigeneity over time. Students in Year 3 in 2009 were in Year 5 in 2011 and in Year 7 in 2013. Similarly, the Year 5 cohort in 2009 was in Year 7 in 2011 and in Year 9 in 2013. The data for the two cohorts by indigenous status are illustrated in Figure 2 (blue/dark lines represent Indigenous student performance).

The data in Figure 2 are revealing. If is very clear that the performance changes over time for the Indigenous and non-Indigenous students in each cohort were virtually identical. One interpretation of these data is that the mathematics learning taking place in mathematics classes across Australia (on average) is not advantaging or disadvantaging one group over the other. On the other hand, schools and mathematics teachers (on average) have not been able to bridge the gap in performance for Indigenous students that first appears at Year 3. One deleterious consequence is that the persistence in Indigenous students' lower achievement levels reinforces the low-achieving stereotyping of this group of Australian students. The issue raised by these data relates to Indigenous students' mathematics learning opportunities prior to the Year 3 level. Does the disadvantage stem from their experiences in the early years of schooling and/or prior to school entry? Alternatively, English language (Literacy) may be implicated. This is examined next.

Exploration 2: Numeracy and Literacy Score Comparisons by Indigenous Status

As mentioned earlier, the sample comprised NAPLAN data for students from 26 schools which were associated with all or part of the *Make it Count* project. The data consisted of the Numeracy and Literacy test scores for students in Years 3, 5, 7, and 9 at these schools for the years 2008-2011.Though strictly anonymised, the files also provided information about each student's Indigeneity, sex, language background, and socio-economic status. The overall sample size for each calendar year was around 4000 students, divided more or less equally among the four grade levels. The percentage of Indigenous students ranged from 7% (in 2008) to just over 9% (in 2011).

The strength (and direction) of a relationship between two variables is often expressed in terms of a correlation coefficient. For our first investigation we computed, for the four years for which we had data and for each grade level, the bivariate correlations (Pearson r) between the NAPLAN Numeracy score and each of the NAPLAN Literacy measures for Indigenous and non-Indigenous students. Space constraints prevent inclusion of the full set of data. Results for NAPLAN 2011 are shown in Table 1 in the next section.

When comparing correlations between different sets of variables it must be recognised that the estimate of a correlation coefficient stabilises with increasing sample size. According to Schönbrodt and Perugini (2013, p. 609) "results indicate that in typical scenarios the sample size should approach 250 for stable estimates". At each grade level, the sample size of non-Indigenous students, but not of each group of Indigenous students, always exceeded 250. To minimize the possibility that differences in the sample sizes of the two groups might confound apparent differences in the strengths of the relationships we used the bootstrapping method, a facility available in current versions of SPSS, to obtain robust values for the correlation coefficients.

> The bootstrapping method originated in the pioneer work of Efron (1979) .... who theorized that one is able to simulate the sample distribution around a statistic (e.g., mean, variance, correlation coefficient, etc.) through the creation of multiple samples with replacement (usually thousands or tens of thousands of runs). Through this method one is able to simulate the population distribution of a statistic (... correlation....) with confidence. (Sideridis & Simos, 2010, p. 118)

Correlations for the NAPLAN 2011 data, using the bootstrap method were also calculated. These are also shown in Table 1. Bootstrap results were based on 1000 bootstrap samples.

Results

The Pearson (bivariate) correlations between the 2011 NAPLAN Numeracy scores and each of the 2011 NAPLAN Literacy scores for the four grade levels at which the test was administered are shown separately for non-Indigenous and Indigenous students in Table 1. Correlations obtained using the bootstrap methods are shown in brackets.

Table 1
Correlations between NAPLAN 2011 Numeracy and each Literacy Score by Year Level and Indigeneity. Bootstrap correlations are shown in brackets

	Correlations* between Numeracy and…			
	Reading	Writing	Spelling	Grammar
Non Indigenous students				
Year 3	.742 (.741)	.550 (.549)	.662 (.665)	.700 (.701)
Year 5	.691 (.688)	.547 (.547)	.576 (.575)	.676 (.676)
Year 7	.672 (.670)	.524 (.524)	.625 (.525)	.661 (.662)
Year 9	.700 (.701)	.438 (.431)	.540 (.539)	.670 (.671)
Indigenous students				
Year 3	.620 (.620)	.452 (.452)	.661 (.661)	.680 (.680)
Year 5	.566 (.566)	.497 (.500)	.551 (.550)	.625 (.621)
Year 7	.704 (.704)	.576 (.575)	.465 (.471)	.731 (.732)
Year 9	.716 (.712)	.503 (.508)	.548 (.560)	.712 (.723)

*Note: All correlation coefficients in Table 1 were statistically significant at p<. 000.

From Table 1 it can be seen that:

- There is much overlap between the values of the correlation coefficients calculated using the traditional (Pearson) method and using the bootstrap method.
- For both groups of students, the correlations between the Numeracy and Reading and between Numeracy and Grammar scores were higher than those between the Numeracy and Writing and Numeracy and Spelling values.
- At each grade level, the highest correlation for non-Indigenous students was consistently between Numeracy and Reading; for the Indigenous students, the highest correlation was between Numeracy and Grammar (using the bootstrap method correlations).

The strength of the relevant correlation coefficients found for the NAPLAN 2008-2010 test data broadly mirrored those found for the 2011 NAPLAN data[3]. Details are not

[3] Between 2008 and 2010 the writing assessment was based on a narrative task. A persuasive task has been used since 2011.

presented because of space limitations. Instead, for each calendar year and the four grade levels at which the NAPLAN test is administered, the Literacy measure found to have the strongest correlation with the Numeracy score is presented in Table 2. To minimize the effect of different sample sizes, the values of the correlation coefficients reflected in the table are those calculated with the bootstrap method.

Table 2
Highest Correlation between NAPLAN Numeracy and Literacy Scores

	Indigenous students				non-Indigenous students			
Gr	2008	2009	2010	2011	2008	2009	2010	2011
3	Reading	Reading	Reading	Grammar	Reading	Grammar	*Reading/ Grammar	Reading
5	Grammar	Grammar	Grammar	Grammar	Reading	Reading	Reading	Reading
7	Grammar	Reading	Reading	Grammar	Reading	Reading	Reading	Reading
9	Reading	Reading	Spelling	Grammar	Reading	Grammar	Reading	Reading

*Note: equal correlation coefficients

From Table 2 it can be seen that:

- With one exception (2010, Year 9, Indigenous students), for both Indigenous and non-Indigenous students, the highest correlations were between Numeracy and Reading or Numeracy and Grammar.
- For non-Indigenous students the strongest relationship tended to be between Numeracy and Reading (all except 2009 Year 3 and 2009 Year 9)
- For Indigenous students the results were more variable. The relationship between Numeracy and Reading was strong for students in Year 3 but for students in Year 5 the highest correlation was consistently between Numeracy and Grammar. There were no consistent patterns for the Year 7 and Year 9 data.

Correlation coefficients *per se* do not enable causal inferences to be drawn. Thus, multiple regression analyses were carried out to determine the impact of the different Literacy measures on student Numeracy scores. Space constraints prevent the inclusion of these results in this paper but the NAPLAN Reading score was regularly found to be the best predictor of the NAPLAN Numeracy score for non-Indigenous students. Consistent with the findings summarised in Table 2, for Indigenous students Grammar was the best predictor of the Numeracy score almost as often as was the Reading score.

Final Comments

In this exploratory study, we examined the cross-sectional and longitudinal trends in national NAPLAN numeracy data for Indigenous and non-Indigenous students. Using a convenience sample of students from schools participating in the *Make It Count* project, we also considered the correlational relationships between NAPLAN numeracy and literacy results to identify patterns of similarity and difference for Indigenous and non-Indigenous students. The findings raise issues related to the identification of potential underlying factors implicated in the relative under-performance of Indigenous students in numeracy.

The patterns in the national NAPLAN numeracy data indicate that the performance of both Indigenous and non-Indigenous students at all grade levels have remained fairly

consistent over time, and that, on average, school experiences of numeracy learning do not appear to have advantaged or disadvantaged either group differently.

The correlational explorations of NAPLAN numeracy and literacy data for the students in the schools participating in the *Make It Count* project revealed that different aspects of literacy are more strongly related to numeracy performance for the two groups. For non-Indigenous students, reading appears to have the strongest relationship to numeracy performance, while for non-Indigenous students, both grammar and reading appear to be implicated.

Based on the findings described above, we suggest that future research should focus on the early years of schooling and pre-schooling experiences of Indigenous children, and that more work is needed on the understanding the relationship between the numeracy and literacy learning of Indigenous students.

References

ACARA. (2012). *National assessment program. Literacy and numeracy [NAPLAN]. National report for 2012*. Retrieved from http://www.nap.edu.au/verve/_resources/NAPLAN_2012_National_Report.pdf

ACARA (2011). NAP. National assessment program. The tests. Retrieved from. http://www.nap.edu.au/naplan/the-tests.html

Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. *Annals of Statistics, 7*, 1-26.

Forgasz, H. J., Leder, G. C., & Halliday, J. (2013). The *Make it Count* project: NAPLAN achievement evaluation. In V. Steinle, L. Ball & C. Bardini (Eds.), *Mathematics education: Yesterday, today and tomorrow* (Proceedings of the 36th annual conference of the Mathematics Education Research Group of Australasia), pp. 298-305. Melbourne, VIC: MERGA.

Messick, S. (1991). Validity of test interpretation and use. In M.C. Alkin (Ed.), *Encyclopedia of educational research* (6th ed , pp. 1–29). New York: Macmillan.

Nichols, S., & Berliner, D. (2007). *Collateral damage. How high stakes testing corrupts America's schools.* Cambridge, Massachusetts: Harvard Education Press.

Popham, W.T. (1997). Consequential validity: Right concern – wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13.

Polesel, J., Dulfer, N., & Turnbull, M. (2012). The experience of education: The impacts of high stake testing on students and their families. University of Western Sydney: The Whitlam Institute. Retrieved from http://www.whitlam.org/_data/assets/pdf_file/0008/276191/High_Stakes_Testing_Literature_Review.pdf

Postlethwaite, T. N., & Kellaghan, T. (2009). National assessment of educational achievement. Paris: UNESCO, International Institute for Educational Planning. Retrieved from http://www.iiep.unesco.org/fileadmin/user_upload/Info_Services_Publications/pdf /2009/EdPol9.pdf

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*(5), 609-612.

Shepard, L. A. (1997). The centrality of test use and consequences for validity. *Educational Measurement: Issues and Practice, 16*(2), 5-8.

Sideridis, G. D., & Simos, P. (2010). What is the actual correlation between expressive and receptive measures of vocabulary? Approximating the sampling distribution of the correlation coefficient using the bootstrapping method. *The International Journal of Educational and Psychological Assessment, 5*, 117-133.

Thomson, S., De Bortoli, L., & Buckley, S. (2013). *Highlights from the full Australian report: PISA 2012: How Australia measures up.* Camberwell: ACER.

Thomson, S., Hillman, K, Wernet, N, Schmid, M., Buckley, S., & Munene, A. (2012). *Monitoring Australian year 4 students' achievement internationally: TIMSS 2011.* Camberwell, Vic, Australia*:* Australian Council for Education Research.