cogent
education

*Corresponding author: R. Nazim Khan,
School of Mathematics and Statistics,
The University of Western Australia,
Crawley, Western Australia
E-mail: nazim.khan@uwa.edu.au

Reviewing editor:
Sammy King Fai Hui, The Education
University of Hong Kong, Hong Kong

Additional information is available at
the end of the article

EDUCATIONAL ASSESSMENT & EVALUATION | RESEARCH ARTICLE

Effect of national partnerships on NAPLAN

Hui Ping Chua[1,2], R. Nazim Khan[1]*, Stephen Humphry[3] and Robert Hassell[4]

Abstract: The annual National Assessment Program—Literacy and Numeracy (NAPLAN) measures the literacy and numeracy skills of primary and secondary students in Australia. Under the three Smarter Schools National Partnership, additional funding is provided to the independent schools with the expectation of improving student performance. Using multilevel modelling to account for within-school variables and demographics, we analyse NAPLAN data from the 2008–2011 tests for a sample of independent schools to estimate the effect of the National Partnerships on student performance. The results indicated that on average male students performed higher in the numeracy test but scored lower in each of the literacy tests. Aboriginal and Torres Strait Islander students scored lower in numeracy, reading, writing, and grammar and punctuation. Students from low socioeconomic status schools performed lower in writing.

Subjects: Educational Change & School Reform; School Leaders & Managers; Secondary Education; Mathematics; English; Teaching & Learning; Gender & Sexuality; Assessment; Curriculum; Education Politics

Keywords: education; indigenous effect; gender effect; literacy; multilevel modelling; NAPLAN; numeracy

1. Introduction

The National Assessment Program—Literacy and Numeracy, better known as NAPLAN, was first administered in 2008 to Australian students in grades 3, 5, 7 and 9. It tests these two foundation skills of learning, using separate tests for numeracy, reading, writing and language conventions, the latter including spelling, and grammar and punctuation. These five tests (numeracy, reading, writing,

ABOUT THE AUTHORS

Chua Hui Ping (BSc, DipEd) is a Mathematics teacher, R. Nazim Khan (PhD, BSc, BE) is a lecturer in Mathematics and Statistics, Stephen Humphry (BPsych, PhD) is a psychometrician and Robert Hassell (BA, DipEd) is an educational consultant, who are interested in assessment matters and strategies for improving teaching and learning. They have authored several papers in this area and presented talks in international conferences in education.

PUBLIC INTEREST STATEMENT

Large-scale testing data are collected in Australia and a range of other countries. However, this source of information is generally underutilised as a way to evaluate the effectiveness of educational programmes and initiatives. This study uses data from Australia's National Assessment Program—Literacy and Numeracy (NAPLAN) to evaluate whether National Partnerships programme had a positive impact on results in Western Australian independent schools. Although we did not find any evidence that the Partnership programme improved student results, several other variables associated with performance were identified, in particular gender and indigenous status. This study illustrates how longitudinal data from large-scale tests can be used in programme evaluation taking into account multiple background factors, such as indigenous status and socioeconomic status, which may be relevant to programme effectiveness.

cogent•oa

spelling, and grammar and punctuation) are taken each year by all students in grades 3, 5, 7 and 9 in the second week of Term two. The items in each of the tests are developed with the advice of specialist officers from departments concerning indigenous education, English as a second language and special needs education (ACARA, 2010). The raw score for each test is converted into a NAPLAN-scaled score out of 1,000. Since different tests are used each year, methods of equating NAPLAN scores between years have been developed, allowing comparisons across years (ACARA, 2010).

The NAPLAN tests have been developed, in part, to provide a common basis to compare educational outcomes between Australian states. The main aim of NAPLAN is to provide a snapshot of the literacy and numeracy skills of Australian students and to inform schools and government in programme and policy development. NAPLAN aids identification of curriculum areas that require more attention from government and schools, and helps identify schools requiring additional support.

Due to the longitudinal nature of NAPLAN testing, the progress of students over time can be measured to some extent. Thus, NAPLAN can allow teachers and schools to identify students in need of extra support as well as those who are excelling in their age group (ACARA, 2011b). Analysis of strengths, weaknesses and effectiveness of teaching methods can also be performed. NAPLAN is useful in setting targets for academic achievement and allows monitoring of national changes in literacy and numeracy standards over time. The progress of indigenous students in literacy and numeracy outcomes is of particular interest.

An aspect of this study is on the effectiveness of special programmes in independent schools aimed at improving literacy and numeracy for their students. Student performance is based on NAPLAN results. Independent schools are defined as those which are not for profit and not administered by the Government of Western Australia, but are registered with the Department of Education Services (https://www.ais.wa.edu.au/independent-schools). The independent schools in this study are members of the Association of Independent Schools of Western Australia (AISWA). These schools also take part in the AISWA Literacy and Numeracy National Partnership (LNNP) under which reforms are made to improve learning in these two key foundation areas (AISWA, n.d.a). Some schools are also in the AISWA Low Socio-Economic Status National Partnership (Low-SES NP) which focuses on students from disadvantaged schools (AISWA, n.d.b). Emphasis is placed on developing and improving evidence-based teaching and strong leadership within the classroom in addition to providing support in school planning and classroom management.

2. Hypotheses
The following hypotheses are of particular interest.

> Hypothesis 1: Indigenous students, defined here as being of Aboriginal or Torres Strait Islander (ATSI) descent, are expected to have lower achievement in each of the NAPLAN tests.

Reduced cultural exposure and knowledge can affect performance in both the literacy tests and the worded numeracy questions (Wigglesworth, Simpson, & Loakes, 2011). Data indicate that the proportions of Australian indigenous students achieving benchmarks in Reading, Writing and Numeracy are significantly lower compared to the nation as a whole (Gray & Beresford, 2008).

> Hypothesis 2: It is expected that students from remote and very remote schools will perform worse in both literacy and numeracy tests. This is again due to reduced cultural exposer, knowledge and resources.

Remote and very remote are defined by a remoteness index, based on the area of the location and its population (Zhao & Guthridge, 2008). Intuitively, they are locations far from any other population centre, have relatively small populations (only a few thousand). About 24.5% of the locations are classified as remote, and about 9.5% as very remote.

Hypothesis 3: Students with a language background other than English are expected to perform worse than English-speaking students.

Previous research has shown that unfamiliarity with the English language can affect performance in word-based mathematics problems (Varughese, 2009, Chapter 2). In particular, translating to and from another language can lead to misinterpretation. Cummings, Kintsch, Reusser and Weimer (1988) reported that mathematics problems with abstract language were more likely to be miscomprehended by non-English speakers. Further, language can appear ambiguous if interpreted in a different context due to unfamiliarity with language.

Hypothesis 4: Students from schools with a higher Socio-Economic Status (SES) or Index of Community Socio-Educational Advantage (ICSEA) are expected to have a higher overall achievement level in the NAPLAN tests.

Considine and Zappalá (2002) identified SES as a key predictor of student academic achievement.

Hypothesis 5: Schools in the Association of Independent Schools of Western Australia Literacy and Numeracy National Partnership or Low-SES National Partnership are expected to have a higher level of improvement on average than those that are not, since these schools are developing programs focusing specifically on literacy, numeracy and general educational improvement.

Hypothesis 6: Gender is expected to be a significant predictor of student performance, with male students hypothesised to perform better in the numeracy test and female students expected to perform better in the literacy tests. (Proud, 2009)

Rothman and McMillan (2003) found that in Australia girls scored higher on reading and comprehension tests, while boys scored higher in mathematics. This trend has also been observed in other countries (OECD, 2014, p. 1).

Hypothesis 7: No difference in improvement is expected to be found between the cohorts of students, that is the 2008–2010 progression group and the 2009–2011 group.

It is reasonable to expect that the cohorts by year will not differ significantly in ability or demographics.

3. Method

3.1. Data collection

The NAPLAN and associated data were provided by the Association of Independent Schools of Western Australia (AISWA) which in turn received the data from the Australian Curriculum Assessment and Reporting Authority (ACARA). The data consist of the Western Australian NAPLAN test results from 2008–2011, student demographic data and several characteristics of the schools. The nature of NAPLAN test administration allows student results to be matched across years, as depicted in Figure 1.

Figure 1. The progression of students from 2008 to 2011.
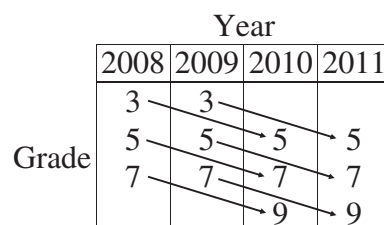
However, since each student is only tested once every two years, matching data will not be available at this stage for students from grade 9 in 2008 or 2009 and from grade 3 in 2010 and 2011. These and any other incomplete records were discarded to avoid issues with missing data. In addition, 8 schools with fewer than 10 students each were removed from the study.

The final data-set consists of 8,266 students from 42 AISWA schools. These students are from one of grades 3, 5 or 7 in 2008 or 2009 who were also tested in 2010 or 2011 in grades 5, 7 or 9, respectively. It was assumed that students did not change schools during the testing period. The gender, ATSI status and Language Background Other Than English (LBOTE) status of each student were also collected during testing. In addition, as reported by Goldstein (1997), students have been found to perform better in high achievement schools, that is, student achievement can be affected by the performance of their peers even after adjustment for their own previous results. Thus, for each grade in each school, the average NAPLAN test scores were calculated and included in modelling (Appendix A).

For each school, their status as a LNNP or Low-SES National Partnership school was provided, in addition to a classification of their location as either metropolitan, provincial, remote or very remote. The SES and ICSEA values for each school were also recorded. ICSEA was specifically developed by the Australian Curriculum Assessment and Reporting Authority to allow for the comparison of NAPLAN scores across schools (ACARA, 2011a). It is a measure of the educational advantage due to the family background of the students and is calculated using student information from the schools, including parental occupation and education level. When this information was unavailable, it was estimated with the use of census data from the Australian Bureau of Statistics. School characteristics such as location, the proportion of indigenous students and the proportion of LBOTE students are also taken into account in ICSEA calculations. Table 1 contains a description of each of the variables.

For each of the five tests, histograms were plotted of the first and second test results to investigate the normality of the data. Figures 2 and 3 indicate that normality is a reasonable assumption.

3.2. Multilevel analysis

Multilevel models (Bliese, 2009; Kahn, 2011), also known as random coefficient models (de Leeuw & Kreft, 1986), hierarchical linear models (Caldas & Bankston III, 1999; Raudenbush & Bryk, 1986), mixed models (Demindeko, 2013) or variance component models (Goldstein, 2011), are used to model hierarchically structured, nested or clustered data (Hox, 2010; Rasbach, Steele, Browne, & Prosser, 2004). Educational data are a prime example with students nested within classes, which are further nested within schools. It is important to take this inherent nested structure into account in the data analysis as ignoring it can lead to incorrect or misleading conclusions. Multilevel models were fitted to the data using the nlme and multilevel packages in R. Due to the five distinct elements of NAPLAN testing, specifically the numeracy, reading, writing, spelling, and grammar and punctuation tests, a separate analysis was conducted to model each of the second test results on the potential predictor variables discussed earlier.

The model involves using level one or student-level predictors and level two or school-level predictors to explain the two sources of variation in the response. In addition, consider, for example, the hypothesised effect of gender on the Numeracy test results. Suppose male students perform better than the female students in the first Numeracy tests (from 2008 or 2009) but both groups of students *improve at the same rate*. If modelling is carried out only on the second tests (from 2010 or 2011) where results from the first test are included as predictors, gender would not be a

| Table 1. The variables collected for use in analysis of the NAPLAN testing results			
Variable name	**Level**	**Type**	**Description**
startyear	Student	Categorical	= 2008; student's first NAPLAN result is from 2008
			= 2009; first result is from 2009
startgrade			= 3; student's first NAPLAN result is for grade 3
			= 5; first result is for grade 5
			= 7; first result is for grade 7
gender	Student	Categorical	= M; student is male
			= F; student is female
atsi	Student	Categorical	= Y; student is indigenous
			= N; student is not indigenous
lbote	Student	Categorical	= Y; student has a non-English language background
			= N; student has an English background
lnnp	School	Categorical	= Y; school is in the Literacy and Numeracy National Partnership program
			= N; school is not in this programme
lowses	School	Categorical	= Y; school is in the Low-SES National Partnership program
			N; school is not in this programme
ses	School	Continuous	The socioeconomic status value for the school
icsea	School	Continuous	The index of community socio-educational advantage for the school
location	School	Categorical	= Metropolitan; school is in the metropolitan area of WA
			= Provincial; in a provincial area
			= Remote; in a remote area
			= Very Remote; in a very remote area
Nfirst, Rfirst, WNfirst, Sfirst, or Gfirst	Student	Continuous	Respectively, the first Numeracy, Reading, Writing, Spelling or Grammar and Punctuation NAPLAN result available for the student (from 2008 or 2009)
Nfirst.mean, Rfirst.mean, WNfirst.mean, Sfirst.mean, or Gfirst.mean	School	Continuous	The school grade average performance for the first Numeracy, Reading, Writing, Spelling or Grammar and Punctuation NAPLAN tests, respectively
Nsecond, Rsecond, WNsecond, Ssecond, or Gsecond	Student	Continuous	The second NAPLAN test results available for the student (2010 or 2011)

significant variable as its effect has already been adjusted for (by including in the model the result of the first test).

Thus, for a more thorough investigation, multilevel models were fitted to each of the first and second NAPLAN tests. For each model fitted, the assumptions of normality, homoscedasticity (unchanging variance at various values of a predictor variable) and linearity were checked using scatterplots and Quantile-Quantile plots of the model residuals.

cogent • education

Figure 2. Histograms for each of the first NAPLAN test results for all students.

Notes: N-Numeracy; R-Reading; WN-Writing; S-Spelling; G-Grammar and punctuation.
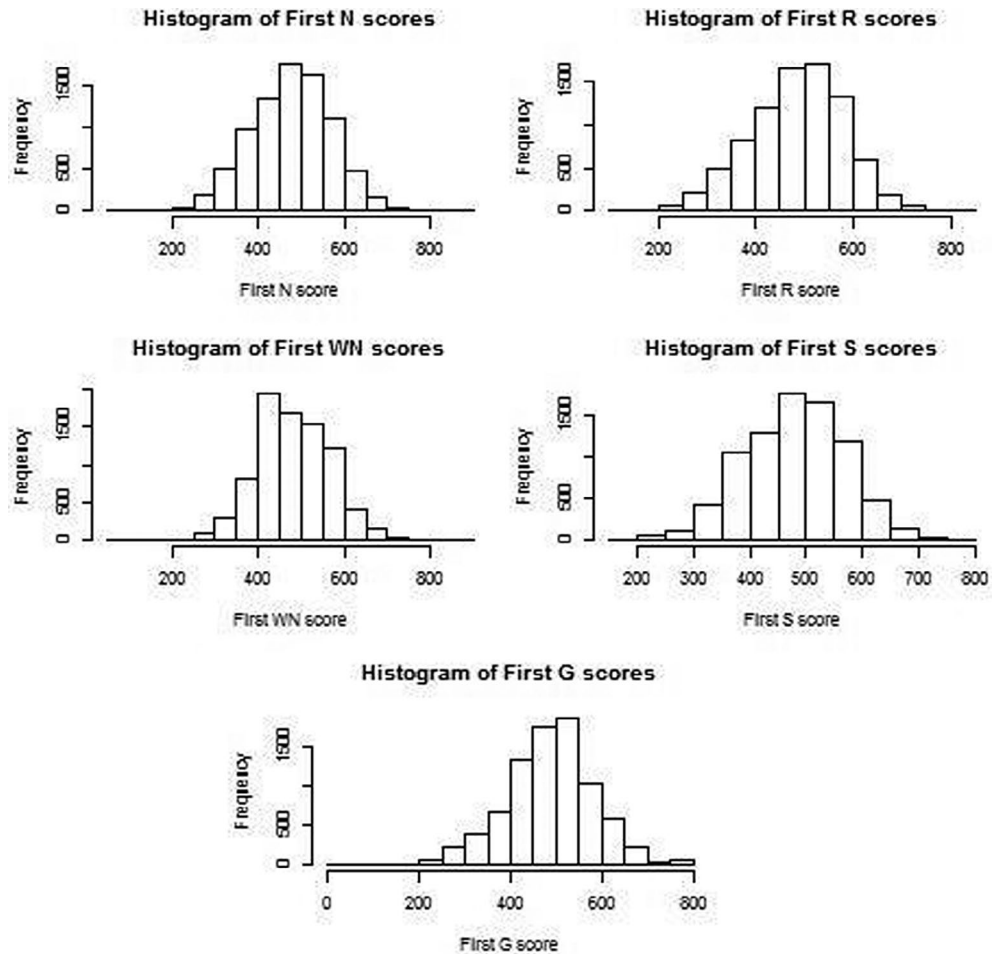
4. Results

As given in Table 1, there are 20 potential predictors for each of the models in addition to the possibility of random coefficients. Initial attempts at modelling using all predictors were obstructed by the failure of the algorithms to converge. To address this, principal component analysis (PCA) was carried out to reduce the number of variables and thus estimators required.

PCA was conducted on the 12 continuous variables; Nfirst, Rfirst, WNfirst, Sfirst, Gfirst, Nfirst.mean, Rfirst.mean, WNfirst.mean, Sfirst.mean, Gfirst.mean, ses and icsea (each defined in Table 1). This resulted in the first three principal components capturing a total of 88.3% of the variability in the data whilst the remaining components each capture less than 2.8% of the variation. The first principal component (pc1) captures 65.2% of the variability, the second (pc2) 13.6% and the third (pc3) 9.4%. These three components are sufficient for use in modelling under an arbitrary 85% cumulative variance stopping rule. Thus, the 12 continuous variables were reduced to three principal components, pc1, pc2 and pc3.

Details of these principal components are given in Table 2. The first component represents a weighted average of the continuous variables, the second represents a contrast of a weighted average of the first score against a weighted average of the SES and ICSEA values whilst the third represents a contrast of a weighted average of the first scores, SES and ICSEA against a weighted average of the mean first scores.

cogent ·· education

Figure 3. Histograms for each of the second NAPLAN test results for all students.

Notes: N-Numeracy; R-Reading; WN-Writing; S-Spelling; G-Grammar and punctuation.
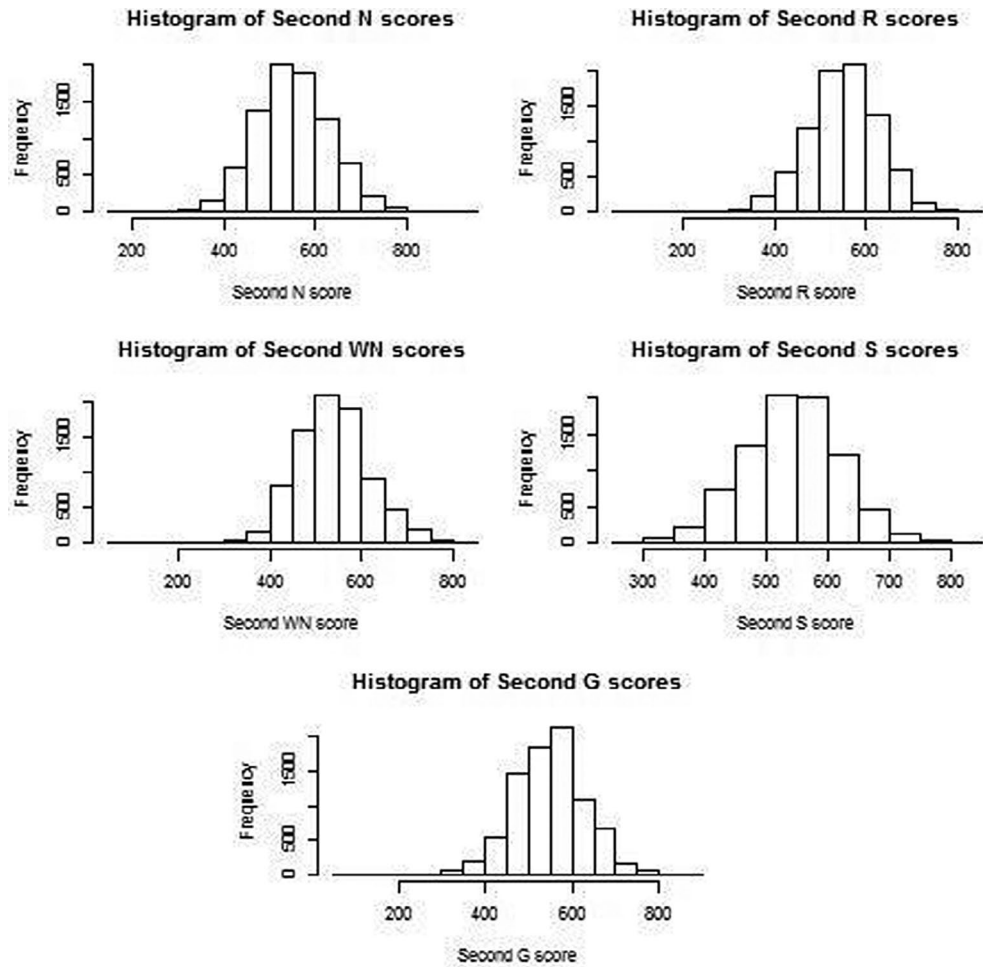
Table 2. Composition of the first three principal components			
Variables	**pc1**	**pc2**	**pc3**
Nfirst	−0.303	−0.108	0.206
Rfirst	−0.289	−0.149	0.363
WNfirst	−0.287	−0.126	0.253
Sfirst	−0.289	−0.208	0.283
Gfirst	−0.281	−0.147	0.407
Nfirst.mean	−0.333		−0.290
Rfirst.mean	−0.336		−0.288
WNfirst.mean	−0.333		−0.286
Sfirst.mean	−0.329		−0.322
Gfirst.mean	−0.333		−0.270
ses	−0.106	0.664	0.244
icsea	−0.120	0.656	0.201

The value of pc1 increases if the continuous variables decrease in value, due to the negative weights. That is, students with an overall low performance in the first tests who come from low performing schools with low SES and low ICSEA will have higher pc1 scores. The second principal component has a higher value if the SES and ICSEA of the student's school is high and the student's first

NAPLAN test results are low. That is, pc2 scores are higher for students who are underperforming in the first tests relative to their school's SES or ICSEA. The third principal component has a higher value if the individual results of the first test and the school's SES and ICSEA are high, and the school-grade mean first test results are lower. This component selects students who have performed above the grade average for a school with given SES and ICSEA values and low pc3 scores can identify students with low performance in the first tests after adjusting for the school-grade mean, and school SES and ICSEA.

In total, ten multilevel models were fitted to the NAPLAN data. The coefficient estimates are summarised in Table 3 where the first and second columns under each NAPLAN test heading refer to the model on the first (2008 or 2009) and second (2010 or 2011) tests, respectively.

5. Discussion

We will discuss the results with reference to the hypotheses stated in the Introduction and then further comment on the implications of the results. It is also expected that relationships could exist between the results from the different NAPLAN tests. For example, a higher reading result could predict a higher numeracy result (Perso, 2011). While our interest is not on the effect of the results of first test on the performance in the second test, this effect needs to be included in the statistical models. However, these effects will not be discussed.

Hypothesis 1: Indigenous students are expected to have lower achievement in each of the NAPLAN tests.

Table 3. Summary of the model coefficients for each of the first and second NAPLAN tests										
	Numeracy		Reading		Writing		Spelling		Grammar and punctuation	
	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd
Intercept	135 (441)	534 (66)	156 (191)	565 (151)	216 (123)	540 (93)	406 (35)	448	69 (493)	557 (55)
ATSI	−18		−32		−19				−39	
Gender M	+16	+26	−12		−25	−19	−13	−4	−21	−8
2009	−0.86		+7	−7 (105)		−7	+5		+20	
Grade 5	+86 (252)	+13	+90	+15 (90)	+77	+14	+89	+21	+105 (443)	−11
Grade 7	+156 (258)	+17	+138	+16 (84)	+129	+26	+141	+14	+133 (298)	
2009*Grade 5	+15	−6			+7			−8	−19	−10
2009*Grade 7				+14		+7		+9		−11
ICSEA	+0.24		+0.24		+0.19				+0.32	
LNNP				−22						
LowSES				−14	−27	−26		−789		−14
Remote					+47					+39
Very Remote							−133	−280		−53
pc1		−19		−17		−15		68		−19 (4)
pc2		−6		−6		−8 (83)		−50		−10
pc3	135	28 (25)	156	34	216	27 (30)	406	−130	69	35
σ^2	4,036	2,191	5,336	1,838	3,872	2,778	4,514	1,271	5,963	2,435

Note: Appropriate, variances of random terms are given in brackets next to the coefficient.

ATSI status was found to have a negative effect on the scores attained for the first numeracy, reading, writing, and grammar and punctuation tests (data from 2008 and 2009 for students in grades 3, 5 and 7), but no significant effect was found for the second tests (from 2010 and 2011 for students in grades 5, 7 and 9, respectively) after adjusting for the first test result. Thus, while indigenous students are at a disadvantage and achieve lower results than non-indigenous students for these aspects, their rate of improvement over the two-year period is found not to differ from that of the non-indigenous students. ATSI status did not affect spelling test results.

Hypothesis 2: Students from remote and very remote schools will perform worse.

Students from remote schools were found to score higher in the first writing test when compared to students from metropolitan, provincial or very remote schools, while no additional effect was found for the second test. That is, after adjusting for the effects of the other factors, students from remote schools score higher in writing but improve at the same rate as those from other schools.

For grammar and punctuation, school location was not found to be a significant factor in the first test results. However, students from remote schools were found to perform higher in the second test while students from very remote schools were found to score lower. This suggests that students from remote schools have a higher rate of improvement than those from metropolitan or provincial schools, but very remote students deteriorated in this area.

It was also found that very remote students performed worse in spelling in the first test, and also deteriorated in the second test.

Hypothesis 3: LBOTE students are expected to perform worse.

The results show that LBOTE status had no significant effect on performance. Thus, NAPLAN seems to accommodate LBOTE students.

Hypothesis 4: Students from a higher SES or ICSEA are expected to perform better.

This hypothesis is related to the principal components. Based on the principal components, an individual performance in the first test that is higher than the school's grade average score results in a predicted increase in each of the second numeracy, reading, writing, and grammar and punctuation test performances. Further, except for reading, this effect was not found to vary between schools. For spelling, students from Low-SES and low Index of Community Socio-Economic Advantage (ICSEA) schools who perform below average in the first test tend to improve in the second test.

Hypothesis 5: Schools in Association of Independent Schools of Western Australia Literacy (AISWA) and Numeracy National Partnership (LNNP) or Low-SES National Partnership (NP) are expected to have a higher level of improvement compared to those that are not.

School participation in LNNP was not found to have any significant effect in predicting the first test scores, after adjusting for the other predictors. The second test scores in reading for LNNP schools were lower, indicating that their performance had deteriorated on average. The writing scores for both tests for Low-SES NP schools were lower, indicating that they had performed worse and also deteriorated over time. For reading, spelling, and grammar and punctuation, Low-SES schools performed just as well in the first test, but deteriorated in the second test.

Hypothesis 6: Male students are expected to perform better in numeracy while females are expected to perform better in literacy.

Gender was found to have a significant effect on both the first and second numeracy scores, with male students performing higher in both tests when compared with their female counterparts. This suggests that on average the male students achieve higher scores in their initial tests and improve

at a higher rate, increasing the gap between male and female students. Female students performed better in all other sections for both the tests except reading for which the difference was only in the first test. This shows that on average female students performed better in these tests, and except in reading also improved at a faster rate than male students. While male students performed worse in the first reading test, their rate of improvement was the same as for female students.

> Hypothesis 7: No difference in improvement is expected to be found between the cohorts of students, that is the 2008 to 2010 progression group and the 2009 to 2011 group.

This hypothesis relates to the principal components, and is discussed below for each test.

5.1. Numeracy
Grade 5 students from the 2009 cohort scored higher in the first test compared to those from the 2008 cohort, but performed worse in the second test after adjusting for the first test results. This suggests that the 2008 Grade 5 cohort improved at a faster rate. No other differences between the cohorts were observed.

5.2. Reading
The grade 3 students from the 2009 cohort achieved higher results in the first reading tests than those from the 2008 cohort, but were found to perform lower in the second reading tests after adjusting for this initial result. This suggests that the 2008 grade 3 students improved at a faster rate than those from 2009. The 2009 cohort of grade 7 students were found to progress at a higher rate than the 2008 grade 7 students.

5.3. Writing
Grade 3 students from the 2008 cohort were found to improve their writing faster than those from the 2009 cohort. Similarly, Grade 7 students from the 2009 cohort progressed faster than their 2008 counterparts.

5.4. Spelling
Grade 3 students from 2009 performed higher than those from the 2008 cohort, but both were found to progress at the same rate. Students from Grades 5 and 7 improved at a faster rate than Grade 3 students. Differences were found in the rate of improvement between the 2008 and 2009 cohorts for both grades 5 and 7. Grade 5 students from 2009 did not improve as rapidly as those from 2008, while grade 7 students from 2009 progressed at a higher rate than those from 2008.

5.5. Grammar and punctuation
Grade 3 students from the 2009 cohort were found to score higher in the first test than those from the 2008 cohort, but both cohorts improved at the same rate. In addition, the grade 5 students from the 2009 cohort scored lower and improved at an even lower rate than those from the 2008 cohort. The grade 7 students from both the 2008 and 2009 cohorts performed higher in the first test but the grade 7 students from 2009 progressed at a slower rate.

5.6. Future work
For improvements in future analysis, it has been recommended (see recommendation 6 at http://www.aph.gov.au/binaries/senate/committee/eet_ctte/naplan/report/b02.pdf) to extend the NAPLAN testing for all students from grades 3 to 10. This would allow for the progression of results for students to be more easily investigated in addition to analysing the long-term effectiveness of various educational policy implementations. More frequent testing would also provide more data, allowing difference in improvement rates between grades, such as those found in the numeracy test, to be

investigated. Caldas and Bankston III found in their 1999 study that differences in the family structure within the school, such as the proportion of single parent families, can have an effect on student achievement above that caused by their own family structure (Caldas & Bankston III, 1999). To investigate this and other similar issues, obtaining more detail on the individual students and the make-up of the schools would be beneficial.

Also covered by Caldas and Bankston III was the issue of racial concentration within the school significantly affecting student academic outcomes (Caldas & Bankston III, 1999). While indigenous status, defined by Aboriginal or Torres Strait Islander descent, was collected for each student in this study, with the increased multiculturalism of Australia it could be of interest to determine if any trends could be found for other ethnic groups. Similarly, rather than the binary language variable LBOTE, the influence of individual language backgrounds could be of interest in future investigations.

From the results of the current analysis, particularly for the spelling test predictions, it would be beneficial if data on more schools were available. In addition, a more balanced selection of schools from the various locations would allow for better, more reliable results. Inclusion of boys-only and girls-only schools in the study could lead to additional discoveries. For example, Proud (2009) found that a higher concentration of female students in a class had a negative effect on English results for the male students.

An initial goal in the analysis of the NAPLAN data was to assess the impact of an instructional leadership programme developed by AISWA. In this programme, schools work together with AISWA to improve the leadership skills for both teachers and principals in the hope that this will result in improved learning. This hypothesis is supported in a study by Heck and Hallinger (2010) who found that distributed leadership was related to improved growth in student. The data regarding instructional leadership was insufficient for this study, with several teachers and principals unable to complete the required surveys in time. In the future, when more results are available and when the programme has been running for longer, perhaps this data can be re-analysed.

6. Conclusion

The most significant observations are the effect of ATSI status and gender. ATSI students performed worse in every test except spelling, and girls performed worse than boys in numeracy. Importantly ATSI status was not significant in models for the second test scores based on the first test scores. Nonetheless, it was found that that first test scores were lower for ATSI students on average. This fact would not have been uncovered simply by an analysis of the second test scores based on the first one. It is clear that the ATSI students do start with lower scores, but there is no difference in their progress rate. Fryer and Levitt (2004) reported a similar gap between black and white students in the first two years of school, and attributed this mainly to the lower quality of schools attended by black students. Therefore, it seems that in order to improve the performance of ATSI students, attention needs to be focused on the earlier years.

The increasing gap in numeracy between girls and boys over the school years is evident. Robinson and Lubienski (2011) found an increasing mathematics gender gap in lower school but not in kindergarten. These findings have strong implications for programmes aimed at reducing or removing the gender gap in numeracy.

cogent • education

Author details
Hui Ping Chua[1,2]
E-mail: hui.chua@education.wa.edu.au
ORCID ID: http://orcid.org/0000-0001-6764-348X
R. Nazim Khan[1]
E-mail: nazim.khan@uwa.edu.au
ORCID ID: http://orcid.org/0000-0003-3349-5006
Stephen Humphry[3]
E-mail: stephen.humphry@uwa.edu.au
Robert Hassell[4]
E-mail: rhassell@ais.wa.edu.au
[1] School of Mathematics and Statistics, The University of Western Australia, Crawley, Western Australia.
[2] Rossmoyne Senior High School, Keith Road, Rossmoyne 6148, Western Australia.
[3] Graduate School of Education, The University of Western Australia, 35 Stirling Highway, Crawley 6009, Western Australia.
[4] Association of Independent Schools of Western Australia, Suite 3/41 Walters Drive, Osborne Park 6017, Western Australia.

References
ACARA. (2010). *Acara's submission to the senate education, employment and workplace relations committee: Inquiry into the administration and reporting of NAPLAN testing.* Retrieved October 24, 2015, from http://www.nap.edu.au/_resources/ACARA_(2010)_-_NAPLAN_senate_inquiry.pdf

ACARA. (2011a). *About ICSEA.* Retrieved September 24, 2015, from http://www.myschool.edu.au

ACARA. (2011b). *Parent/carer support.* Retrieved September 28, 2012, from http://www.nap.edu.au/NAPLAN/Parent_Carer_support/index.html

AISWA. (n.d.a). *Literacy and numeracy.* Retrieved September 24, 2015, from http://www.ais.wa.edu.au/national-partnerships/literacy-and-numeracy/

AISWA. (n.d.b). *Low ses.* Retrieved September 24, 2015, from http://www.ais.wa.edu.au/national-partnerships/low-ses/

Bliese, P. (2009). *Multilevel modeling in R (2.3): A brief introduction to R, the multilevel package and the nlme package.* Retrieved September 24, 2015, from http://cran.r-project.org/doc/contrib/Bliese_Multilevel.pdf

Caldas, S. J., & Bankston, C. L., III. (1999). Multilevel examination of student, school, and district-level effects on academic achievement. *The Journal of Educational Research, 93,* 91–100. http://dx.doi.org/10.1080/00220679909597633

Considine, G., & Zappalá, G. (2002) Factors influencing the educational performance of students from disadvantaged backgrounds. In T. Eardly & B. Bradbury (Eds.), *Competing Visions: Refereed Proceedings of the National Social Policy Conference 2001* (pp. 91–107). Sydney: Social Policy Research Centre, University of New South Wales.

Cummings, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20,* 405–438.

de Leeuw, J., & Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics, 11,* 57–85.

Demindeko, E. (2013). *Mixed models: Theory and applications with R* (2nd ed.). Hoboken, NJ: John Wiley and Sons.

Fryer, R. G., & Levitt, S. D. (2004). Understanding the black-white test score gap in the first two years of school. *Review of Economics and Statistics, 86,* 447–464. http://dx.doi.org/10.1162/003465304323031049

Goldstein, H. (1997). Models in school effectiveness research. *School Effectiveness and School Improvement, 8,* 369–395. http://dx.doi.org/10.1080/0924345970080401

Goldstein, H. (2011). Multilevel statistical models In *Wiley series in probability and statistics* (4th ed.). Hoboken, NJ: John Wiley & Sons.

Gray, J., & Beresford, Q. (2008). A 'formidable challenge': Australia's quest for equity in indigenous education. *Australian Journal of Education, 52,* 197–223. http://dx.doi.org/10.1177/000494410805200207

Heck, R. H., & Hallinger, P. (2010). Testing a longitudinal model of distributed leadership effects on school improvement. *The Leadership Quarterly, 21,* 867–885. http://dx.doi.org/10.1016/j.leaqua.2010.07.013

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications. Quantitative methodology series.* London: Routledge.

Kahn, J. H. (2011). Multilevel modeling: Overview and applications to research in counseling psychology. *Journal of Counseling Psychology, 58,* 257–271. http://dx.doi.org/10.1037/a0022680

OECD. (2014). *PISA. Are boys and girls equally prepared for life?* Retrieved from http://www.oecd.org/pisa/pisaproducts/PIF-2014-gender-international-version.pdf

Perso, T. (2011). Assessing numeracy and naplan. *The Australian Mathematics Teacher, 67,* 32–35.

Proud, S. (2009) Girl power? An analysis of peer effects using exogenous changes in the gender make-up of the peer group. Bristol: The Centre for Market and Public Organisation 08/186, Department of Economics, University of Bristol. Retrieved September 24, 2015, from http://ideas.repec.org/p/bri/cmpowp/08-186.html

Rasbach, J., Steele, F., Browne, W., & Prosser, B. (2004). *A user's guide to MlwiN version 2.0.* Bristol: Centre for Multilevel Modelling, University of London Institute of Education.

Raudenbush, S., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59,* 1–17. http://dx.doi.org/10.2307/2112482

Robinson, J. P., & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal, 48,* 268–302. http://dx.doi.org/10.3102/0002831210372249

Rothman, S., & McMillan, J. (2003). *Influences on achievement in literacy and numeracy* (Research Report 36). Sydney: Australian Council for Education Research.

Varughese, N. A. (2009). *Language difficulties in mathematics courses for students from non-English speaking backgrounds in the transition from secondary to tertiary education* (PhD Thesis). Royal Melbourne Institute of Technology, Melbourne.

Wigglesworth, G., Simpson, J., & Loakes, D. (2011). Naplan language assessments for Indigenous children in remote communities. *Australian Review of Applied Linguistics, 34,* 320–343. http://dx.doi.org/10.1075/aral

Zhao, Y., & Guthridge, S. (2008). Rethinking remoteness: A simple and objective approach. *Geographical Research, 46,* 413–420. doi:10.1111/j.1745-5871.2008.00534.x

Appendix A

Let y_{ij} represent the score from the 2010 test and let x_{ij} represent the score from the 2008 test for student i from school j, $i = 1, 2, …, 8,266$, $j = 1, 2, …, 42$. The general intercepts- and slopes-as-outcomes model (Bliese, 2009; Kahn, 2011) with one student level (level one) predictor x_{ij} (the student's score in the 2008 test) and one school level (level two) predictor W_j (the school average NAPLAN score for the 2010 test), is given in Equation (1),

$$y_{ij} = \beta_{0j} + \beta_1 x_{ij} + r_{ij}, \tag{1}$$

where $\quad \beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j}, \tag{2}$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j}, \tag{3}$$

$$\mathrm{Var}(r_{ij}) = \sigma^2,$$

$$\mathrm{Var}(u_{0j}) = \sigma_{u0}^2,$$

$$\mathrm{Var}(u_{1j}) = \sigma_{u1}^2.$$

The error terms r_{ij}, u_{0j} and u_{1j} are assumed to be independently normally distributed with mean 0. They represent respectively the overall error and the school-based error terms for the random intercept and random coefficient. The terms $\sigma_{u0}^2, \sigma_{u1}^2$ and σ^2 are also called *variance components*, as they model the different sources (components) of variation in the model. Substitution of the intercept and slope expressions, (2) and (3), into the model Equation (1) results in the model representation

$$y_{ij} = \gamma_{00} + \gamma_{01} W_j + \gamma_{10} x_{ij} + \gamma_{11} W_j \times x_{ij} + \left(u_{0j} + u_{1j} x_{ij} + r_{ij} \right). \tag{4}$$

From this, it can be seen explicitly that a cross-level interaction occurs between W_j and x_{ij}. In this sense, W_j enters the model as a *moderator* variable, as its value can moderate the effect x_{ij} has on y_{ij}. The term $(u_{0j} + u_{1j} x_{ij} + r_{ij})$ in Equation (4) represents the random effect terms of the model. Again, when more than one student-level or school-level predictor is fitted, interactions between these variables may be added to the model.

Interpretation of the estimated intercept γ_{00} and the slope γ_{10} is aided by the construction of confidence intervals given in Equation (5),

Intercept: 95% CI for $\gamma_{00} = \hat{\gamma}_{00} \pm z_{0.025} \sqrt{\hat{\sigma}_{u0}^2}. \tag{5}$

Slope: 95% CI for $\gamma_{10} = \hat{\gamma}_{10} \pm z_{0.025} \sqrt{\hat{\sigma}_{u1}^2}.$

Interpretation and construction of confidence intervals for the slope term β_{1j} is more complicated due to the possibility of a cross-level interaction term. However, if the interaction term is considered separately, the confidence interval can be constructed as in Equation (5).

Consider for example, the model for the second grammar and punctuation test, given in (6) below, where the subscript ij represents student i in school j. In this model, both the intercept and the coefficient of the first principal component have random variation across the different schools, whilst all other coefficients are fixed.

In Equation (6) u_{0j} and u_{1j} represent the random elements of the intercept and $pc1$, respectively.

$$\text{Gsecond}_{ij} = \beta_{0j} + \beta_{1j}\, pc1_{ij} - 10\, pc2_{ij} + 35\, pc3_{ij} - 8\, sexM_{ij} \tag{6}$$

$$-11\, \text{startgrade5}_{ij} - 10\, \text{startyear2009} * \text{startgrade5}_{ij}$$

$-11 \text{ startyear2009} * \text{startgrade7}_{ij} - 13.56 \text{ lowses}Y_j$

$+ 39 \text{ locationRemote}_j - 53 \text{ locationVeryRemote}_j + r_{ij},$

where $\beta_{0j} = \hat{\gamma}_{00} + u_{0j} = 557 + u_{0j},$

$\beta_{1j} = \hat{\gamma}_{10} + u_{1j} = -19 + u_{1j},$

$\text{Var}(r_{ij}) = \hat{\sigma}^2 = 2435,$

$\text{Var}(u_{0j}) = \hat{\sigma}^2_{u0} = 55,$

$\text{Var}(u_{1j}) = \hat{\sigma}^2_{u4} = 4.$

These estimates can be used to construct approximate 95% confidence intervals (CI) for the coefficients with random variation as given in Equation (7), where $z_{0.025} = 1.96$ the 0.025 critical value of the standard normal distribution is.

Intercept: $95\% \text{ CI} = \hat{\gamma}_{00} \pm z_{0.025} \sqrt{\hat{\sigma}^2_{u0}}$ (7)

$= 557 \pm 1.96 \times \sqrt{55}$

$= (542, \ 572).$

pc1: $95\% \text{ CI} = \hat{\gamma}_{10} \pm z_{0.025} \sqrt{\hat{\sigma}^2_{u1}}$

$= -19 \pm 1.96 \times \sqrt{4}$

$= (-23, \ -15).$

Equations for the other models can similarly be expressed using the coefficients in Table 3.

Cogent Education **(ISSN: 2331-186X) is published by Cogent OA, part of Taylor & Francis Group.**

Publishing with Cogent OA ensures:

- Immediate, universal access to your article on publication
- High visibility and discoverability via the Cogent OA website as well as Taylor & Francis Online
- Download and citation statistics for your article
- Rapid online publication
- Input from, and dialog with, expert editors and editorial boards
- Retention of full copyright of your article
- Guaranteed legacy preservation of your article
- Discounts and waivers for authors in developing regions

Submit your manuscript to a Cogent OA journal at www.CogentOA.com