Interpreting NAPLAN Results for the Layperson

Margaret Wu

University of Melbourne

Email address: m.wu@unimelb.edu.au

Phone: +61 3 8344 0963

Fax: +61 3 9348 2753

Postal address:

Assessment Research Centre,
234 Queensberry Street,
Melbourne Graduate School of Education,
University of Melbourne,
Victoria,
Australia, 3010

Keywords: achievement test, error margin, gain score, confidence level

The Australian federal government's education transparency agenda should begin with providing the layperson with clear guidelines for interpreting the National Assessment Program – Literacy and Numeracy (NAPLAN) results. In particular, the accuracies and limitations of NAPLAN results should be made clear in plain language, so that all stakeholders can make use of NAPLAN results in an informative way.

This document is for the purpose of explaining to the layperson the valid use of NAPLAN results. A technical appendix has been included to enable those with a technical background to check how the conclusions were arrived at.

Fluctuations in test scores

In NAPLAN, for each subject area (numeracy, reading, spelling, grammar & punctuation and writing), there is very limited testing carried out on each student. For example, for Year 5 numeracy, each child is tested on just 40 questions. If David obtained 25 out of 40 on the 2009 test, and Tina obtained 23 out of 40, we cannot make the definitive conclusion that David is better than Tina in numeracy in general. This is because there are many possible questions about year 5 numeracy that could be asked but there is only room on the test for a sample of 40 questions. This means from test to test an individual's score will naturally vary. For example, if we had given David and Tina the 2008 NAPLAN Year 5 numeracy test, it is quite conceivable that David could obtain 23 out of 40, and Tina could obtain 26 out of 40, so we could arrive at the conclusion that Tina is better than David.

So how much should we expect David's scores to vary if tests similar to the NAPLAN Year 5 numeracy test are administered? For a 40-question test, David's scores might vary by as much as ±5 score points (see technical appendix, Note 1). That is, if David obtained 25 out of 40, we expect[1] that his score will range between 20 and 30 should similar tests of 40 questions be given. In percentage terms, the test scores are expected to vary by around ±12%. That is, for a 40-question test, if a student's score is 70% on the test, we expect the range of this student's scores on similar tests to be between 58% and 82%. That is quite a wide range! On the other hand, this should not be surprising to teachers and students, as we all know that our test scores fluctuate from test to test.

Interpreting NAPLAN score for an individual student

Teachers and parents should be aware that a student's NAPLAN score on a test could fluctuate by about ±12%. Consequently, any use of an individual student's NAPLAN result should take this uncertainty into account. Remember that NAPLAN results are based on just one single test of limited test length. A sample of 40 questions is not sufficient to establish, with confidence, the exact numeracy proficiency of a student. The same caution applies to all subject areas tested.

Estimating students' growth

Value-added analysis is the term commonly used for the analysis associated with estimating students' growth. Such an analysis typically involves the computation of gain scores – the

[1] This is based on "95% confidence interval". That is, had many similar 40-question tests been administered, we would expect that 95% of the time, David's score will be between 20 and 30.

difference between the test scores for an individual on two different testing occasions[2]. For example, if David obtained 25 out of 40 this year, and he obtained 30 out of 40 on a test of similar difficulty next year, then the gain score is 5. If Tina obtained 23 this year, and 30 next year, then the gain score is 7. We then compare whether Tina made a greater gain score than David. While it is true that Tina made more gain score than David (7 versus 5), are we confident that Tina actually acquired more numeracy knowledge in the one year period than David, given the fluctuations in test scores as I have shown in the previous section?

How much should we expect a gain score to vary due to random fluctuations of test scores? For comparing the difference in scores on two 40-question tests, the gain score will be expected to vary by around ±7 score points[3]. That is, in the case of Tina, her gain score is expected to range between 0 and 14 score points, if similar tests were administered one year apart. In NAPLAN, a typical year's growth is around 5 score points for a 40-question test. For example, if Year 3 average is 25 out of 40, then Year 4 average will be around 30 out of 40 on the same test. This means that, based on the results of two tests administered one year apart, Tina could show no growth at all, or three years of growth, just through random fluctuation of test scores.

Interpreting growth measures at the individual student level

For an individual student, the growth measures based on two 40-question tests have an error margin greater than one year's growth.

School level results

Student results are aggregated to provide school level results. The variation in the average class score due to the inaccuracies in measuring each student is smaller than at the individual student level. This is because the inaccuracies of each student's test score are averaged across a number of students. However, another source of error becomes important. This source of error comes from the sampling of students. If we ever infer student results to teacher performance, then we need to take into account that each year's cohort of students for a teacher varies. For one year, the teacher may have relatively lower achievers than for another year. The year-to-year, and class-to-class variation in student proficiency levels is quite large. One year there may be a few more high achievers, and another year there may be a few more stragglers.

If Ms Smith's class achieved an average score of 25 out of 40 for numeracy this year, we can expect that the class average for Ms Smith from year to year to vary mostly between 23 and 27, taking into account the variation of student cohort and the inaccuracies in measuring each student[4] (See technical appendix, Note 2). The range of 4 score points (23 to 27) is 10% of the total score, and it is about one year's growth. That is, even if Ms Smith didn't change the way she taught, she could expect her class average score to fluctuate within a range of one year's growth.

[2] In the NAPLAN context, the scores used are not raw scores. A transformation of the raw score is carried out to arrive at an IRT (item response theory) score. For simplicity, I will not go into IRT here. There is no great difference between using raw scores and IRT scores for illustrating error margins.

[3] I have not taken into account equating error here. Equating error is the error involved in placing the current year's test results on the same scale as previous years' results. We have seen in 2009 the equating errors are absolutely huge!

[4] This is a conservative estimate. See technical appendix Note 2.

Interpreting class average for a NAPLAN test

A teacher can expect his/her class average score on a NAPLAN test to vary by around 10% from year to year due to random fluctuations of student cohort and inaccuracies in test scores. If we use the class average to judge a teacher's performance, we need to keep in mind that the class average could be higher or lower to some extent depending on a teacher's "luck" of whether the current cohort of students are relatively better or poorer academically.

Linking student performance to school performance

If we use the term "school performance" to mean the effectiveness of school staff (including the principal and teachers), then linking low student NAPLAN results to low school performance is a conjecture. Statistics can provide numbers in terms of average scores and the spread of the scores, but statistics cannot provide substantive interpretations of why student results are low. Such interpretations are made by people. Any explanation offered by people is a conjecture, a speculation. This is the case whether schools are compared within like-school groups or not. This is the case whether it is the comparison of current student measures or the comparison of growth measures. That is, finding statistically significant differences in student results does not necessarily imply differences in school performance (teacher/principal performance).

A low result for a school could be due to mis-management or teacher ineffectiveness. But equally, it could be due to circumstances unrelated to school performance. For example, there might have been an outbreak of chicken pox in the school and many students were away for lengthy periods just prior to the test. There could be a high number of new migrant students who had difficulties with the English language. Students with NESB status have varying degrees of language proficiency that may not be captured accurately by the like-school variables.

Another reason for calling for caution in linking student performance to school performance is that students' knowledge and skills are cumulative, acquired over the years of students' school life from Year 1, if not earlier. A teacher would typically have taught a student for three months when the NAPLAN tests take place. What a student can do in numeracy, for example, is the knowledge the student acquired over past years of schooling, with some additional knowledge in the three months prior to the NAPLAN test. To attribute students' test scores entirely to the current teacher's performance may not be justified. In relation to Year 7 and Year 9 NAPLAN tests, it is conceivable that some students did not have a solid foundation in primary years and their low scores were not the result of poor teaching by their current teacher or the poor performance of their current school. While we recognise that the blame game will probably not lead us anywhere, we should at least acknowledge that it is difficult to pinpoint the reasons for a student's low performance.

If you are a teacher or a principal, and anyone alleges that you have not performed your work effectively based on NAPLAN student scores, you have a good case to challenge the accuser(s). NAPLAN results simply do not provide unequivocal evidence that someone has not done their work effectively, even when the results are presented together with other school characteristics.

An appropriate way to use school level student results is for education authorities to identify schools with low scores and, in private and in consultation with the school, conduct further investigations into the reasons for low scores. When possible reasons are identified, remedial actions can be taken. Publication of school results without appropriate cautions can only lead to mis-interpretation and mis-information, and in some cases, defamation of school staff.

School comparisons

NAPLAN results alone CANNOT show, with confidence, which schools are more effective and which schools are less effective. Even taking into account of school contextual information such as school socio-economic status, staff numbers and funding breakdowns, we still cannot positively identify poor school performance. This is because school contextual information cannot capture ***all*** *factors that have an impact on student performance* ***other than*** *school performance*[5]*. NAPLAN results and school contextual information provide only indications for further investigation to find more direct evidence of school performance.*

Comparisons at the Jurisdiction Level

In most national assessment programs (e.g., PISA, NAPL-SL, NAPLAN), ACT has come out as the top performing jurisdiction, and NT has come out as the lowest performing jurisdiction[6]. Yet few people attribute these results to the best education system and the best teachers in ACT, and the worst education system and the worst teachers in NT. Most people would attribute the high performance of ACT to the demographic composition of ACT: that there are relatively more public servants in ACT than in other jurisdictions. The low performance of NT can be largely accounted for by the relatively high proportion of students in remote regions. I am not suggesting that students from remote regions should naturally perform lower than those from urban regions. But these are issues that go beyond the best teaching practice and efforts of principals and school management.

So when Queensland and Western Australia performed lower than all jurisdictions except for NT, should we not take into account that the demographic compositions of QLD and WA are not quite the same as those of NSW and VIC? There are also comparability issues such as the age of students at time of testing. For example, in QLD, the average age of students is a little younger than for other jurisdictions. I would not immediately jump to the conclusion that QLD and WA educational authorities have not done the best they could, or that the teachers and schools in these two states have not been as effective as those in NSW, VIC and ACT.

This means that the claim that NAPLAN provides “nationally comparable data” is flawed. While it is true that students did the same tests nationally, it is not necessarily true that the data are directly comparable nationally. The likely washback effect at the jurisdiction level is that QLD and WA rush to rectify their student performance, while ACT, NSW and VIC become complacent about their performance. All of these are likely to be based on misguided interpretations of NAPLAN results.

Summary

Whenever comparative results are presented, always ask the question whether the differences in scores are likely due to random fluctuation or due to real differences. Never accept any

[5] For further reading, see Briggs, D., & Wiley, E. (2008). Causes and effects. In Ryan, K. & Shepard, L (2008) (Eds.). *The Future of Test-Based Educational Accountability*. Routledge, NY. Other chapters in this book are also very informative. Briggs and Wiley stressed that estimates from value-added analysis were not suitable as the primary basis for high-stakes estimates of school effects.

[6] In fact, if they didn’t, it’s a sign that something went wrong in the survey methodology.

comparison of figures if the confidence level of the results is not revealed. Small schools should be particularly vigilant as the natural variation in scores is typically large for these schools.

Above all, remember that NAPLAN results are based on one test of 40 questions administered once a year for each subject area. Your use of NAPLAN results should be based on the confidence level associated with such a test.

From my point of view, the publication of NAPLAN results at the school level will do great harm to Australian education because of the complexities of the interpretations of the results. I hope this paper provides the layperson with some clarification. I have simplified the discussions on some issues, but the messages conveyed should still be valid.

Additional Resources

For other resources on the topic of large-scale assessments such as NAPLAN, I particularly recommend the book by Daniel Koretz: Measuring Up (2008, Harvard University Press). In this book, Koretz explains the issues related to educational testing in plain language without technical jargon. You can read the first chapter at the following link (place cursor on the book and select First Pages):

http://www.amazon.com/Measuring-Up-Educational-Testing-Really/dp/0674028058

Another book that is pertinent to national testing is the book edited by Katherine Ryan and Lorrie Shepard: The future of test-based educational accountability (2008, Routledge). This book was published in the wake of the No Child Left Behind Act, 2001, in the United States, when the government stipulated the requirement of annual assessments at the state level. You can imagine that the issues that are currently debated in Australia are discussed 50 times more widely in the United States. Instead of re-inventing the wheel in Australia, we can gain a great deal of the experiences in the United States over the past eight years. This book is not overly technical. It is suitable for policy makers, administers as well as education researchers. You can find sample pages from Google Books.

Technical Appendix

1. Computation of standard error of measurement.

For a 40-question test, I used a test reliability value of 0.86, and a standard deviation of 7 for the raw score on the test. These figures were obtained from a state-wide testing program for Year 5 numeracy. The standard error of measurement is computed using the formula

$$\sqrt{\left(1-\textit{reliability}\right)}\times\text{standard deviation of test score}$$
$$=\sqrt{\left(\text{1-0.86}\right)}\times 7$$
$$=2.6$$

The 95% confidence interval is computed as $\pm 2\times 2.6=\pm 5.2$

2. Computation of standard error of class mean based on sampling error and measurement error

I assumed an average class size of 25. Further, I assumed the standard deviation of (observed) test scores for one class to be 5 score points. For example, on a 40-question test, the majority of the class test scores could range between 15 and 35.

The standard error due to the sampling of student and measurement error is $\frac{5}{\sqrt{25}}=1$

The 95% confidence interval is computed as $\pm 2\times 1=\pm 2$.

This is a conservative estimate, as I am assuming that a standard deviation of 5 score points includes both measurement and sampling error. Actually measurement error alone is 2.6 score points for each student.