Kevin Wheldall

The Summary Report of the results for the 2012 NAPLAN (National Assessment Program – Literacy and Numeracy) testing was released last month (http://tinyurl.com/99jrreg), focusing on achievement in reading, writing, language conventions (spelling, grammar and punctuation), and numeracy. In brief, the results appear to show little or no change from last year and, more surprisingly, little change compared with 2008 when the first NAPLAN testing was carried out (the ‘base NAPLAN year’).

Because I am currently particularly interested in literacy learning in the early school years, I’ll focus mainly on the results for reading in Year 3. I’ll also focus primarily on the results for New South Wales (NSW) because that is the state in which I live. The interesting comparison data are reported from page 28. Moreover, statistical significance testing of differences is also provided, comparing 2008 with 2012 results and also comparing 2011 with 2012. Using the available information, we can also calculate rough effect sizes (which are not given) from the information provided. (Note that these can only be approximations because we do not have access to the raw scores for each child.) When we are dealing with huge sample sizes, as we are here (close to the whole population, in fact), even very small, trivial differences can show up as being statistically significant. By calculating the effect size, however, we can gain an indication of whether the differences are worth bothering about or not.

If we look, for example, at the achievement scores for reading for Year 3 in NSW for 2008 and 2012, the mean (average) score rose from 412.3 to 425.7, an increase of 13.4 points, which was statistically significant. The rough effect size, however, was about 0.17 at best, which is very small; so small, in fact, that it is regarded as barely having an effect at all. Researchers tend to regard an effect size of 0.2 as being the lowest value to count even as ‘small’. To put this in perspective, John Hattie, in his book Visible Learning, argues that an effect size of 0.4 is what he calls the ‘hinge’ value, meaning that this is the point when interventions become worthwhile.

The results Australia-wide for Year 3 reading were similar with an effect size of about 0.22. Queensland recorded the biggest improvement with an effect size of about 0.44, above Hattie’s hinge and well worth the effort. In this case, the effort is likely to have been the additional year of schooling introduced in Queensland when they introduced a prep/kindy year in 2008. 2011 was the first year that Queensland children in Year 3 were in their fourth year of schooling, when they took the test, as they were in the other states. (The difference between 2011 and 2012 results, although significant, is negligible with an effect size of 0.1, confirming that the big increase was between the years 2008 and 2011.)

The results for reading in Year 5 show very little change occurring over the years. Although some differences are statistically significant, the rough effect sizes are very small.

So what does this tell us in broad terms? It tells us that generally there has been no major improvement in reading performance over the years 2008 to 2012 for children in Years 3 and 5. There is one important proviso, however, and that is the assumption that the annual tests are truly comparable from year to year, as we are assured is the case. If they were found to be not truly comparable, we could draw no meaningful conclusions at all.

Emeritus Professor Kevin Wheldall AM is Chairman of MultiLit Pty Ltd and Director of the MultiLit Research Unit. You can follow him on Twitter (@KevinWheldall) where he comments on reading and education (and anything else that takes his fancy). He also has a blog ‘Notes from Harefield: Reflections by Kevin Wheldall on reading, books, education, family, and life in general’: http://kevinwheldall.com.au.

Latest NAPLAN results: No change

When we are dealing with huge sample sizes, as we are here (close to the whole population, in fact), even very small, trivial differences can show up as being statistically significant.