Learning lessons: What No Child Left Behind can teach us about literacy, testing and accountability

Allan Luke and Annette Woods

Literacy education remains a contentious policy and pedagogic issue for communities, schools, systems, teachers and students – and politicians. Over the past three to four years in Australia, we’ve seen the latest literacy crisis played out in the pages of our newspapers and television current affairs shows. This crisis has predictably led to policy and curriculum initiatives offering simplistic solutions to the latest perceived problems. Predictable? These Australian media and policy responses parallel the debates in the United States over the No Child Left Behind Act (2001). So are there lessons to be learnt for Australian teachers and policy makers in the No Child Left Behind legislation? Here we lay out a brief introduction to the NCLB legislation and its policy effects. We document the official results and the critiques. Our case is that if Australian policy-makers wish to follow the parameters of US policy, they must consider the real problems and risks that have emerged in the past five years. In these brief comments – we put aside larger ideological and political debates.

Our aim is to lay out what the research literature says about the effects of this approach. Some readers will find this paper dry and boring – fair enough. But we here want to offer a scientific and bibliographical resource for teachers who wish to engage with the debate.

What is No Child Left Behind?

The No Child Left Behind Act (2001) (NCLB) began from a focus on improved quality of early literacy acquisition. It was launched as a bipartisan move for US schooling to address criticisms of lagging literacy standards and a growing achievement gap for students from diverse backgrounds. A significant assumption is: that a program of teaching basic literacy code breaking skills and regular testing can generate more equitable results and better achievement by students from cultural and linguistic minority and lower socio-economic backgrounds. NCLB is based on four basic principles: increased accountability; increased flexibility and local control; expanded options for parents; and teaching methods based on a “gold standard evidence base” (www.ed.gov/programs/readingfirst/legislation.html). The legislation sets out a regime of accountability as standardised testing and public reporting, with a drive toward consistency across system and school contexts. What will count as evidence-based pedagogy is set by mandating scripted, highly prescriptive reading curriculum programs and methods, approved by the Federal government.

Here we leave aside for the moment the ongoing conflict-of-interest controversies about the US Federal government selection of programs for funding (Grunwald, 2006)—a matter of ongoing contention in Congressional hearings.
Compliance in relation to teaching methods and testing measures is controlled through sanctions and incentives provided for districts, schools, and teachers. Parental school choice is set up as a way to let the market drive the system: provide the facts about individual school performance and the market will decide. The facts in this case are provided through published league tables of school test performance. There is, of course, the assumption that access to choice can be established as a level playing field for all, with voucher systems in place in some states for families from traditionally disadvantaged communities. To those who have been following the Australian policy debates – these propositions would be familiar familiar.

Key early criticisms of the legislation were based on claims that only a very limited selection of research was taken into account in the National Reading Panel (2000) findings that formed the basic foundation for the NCLB initiatives. Yatvin, Weaver and Garan (2003) claim that the Reading Panel Report is misleading when it claims that 100,000 instances of research were considered. In fact this number represents the estimated number of research papers and reports published in the 30 year period considered, with a much smaller number of research papers actually being judged to fit the ‘gold standard’ of evidenced-based research set up a priori to the research review. The effect of this narrow definition of evidence-based research was the omission of much of the long standing educational research base – including all correlation studies and those that related to single case sites. In fact very small numbers of studies from the possible thousands were included in the recommendations and findings of the Panel. As an example the claims about the effectiveness of phonic instruction in early reading were made on the basis of just 38 studies, with key contradictory research excluded from the review.

Further, the National Reading Panel and subsequent NCLB policy has been critiqued for: discrepancies between the findings of sub group reports and the final reported findings; the omission of findings related to English language learners (ELLs) and the invisibility of special needs issues; incomparability of results across different populations being ignored as findings of a variety of research studies were conflated; and errors in the predictive validity of early reading achievement assessments (see as examples Allington, 2002; Coles, 2003; Garan, 2002; Yatvin, Weaver & Garan, 2003) as examples of this large research base). In a broader critique, Gee (2000) linked the panel’s approach to science with its failure to engage with well-documented new cultures, technologies, and practices of literacy.

The official results and first round of reanalysis

Three years after its inception, the first official results of the policy initiative’s effectiveness were published (Centre on Education Policy (CEP), 2003). In what are effectively state-by-state self-reports, 73% of states and 72% of districts reported improvement on state tests over the first three years of NCLB. Further, 21 states reported a narrowing of the Hispanic achievement
gap and 18 states similarly reported a narrowing of the African American achievement gap. These figures suggest improvement, although not of the scale and mass necessary to reach targets of 100% of students at grade proficiency by 2014. Issues related to equitable distribution of trained teachers across all schools, a lack of support for English Language Learning (ELL) and special needs students, a narrowing of the curriculum and issues of content and construct validity and confidence interval problems with state testing and reporting on some counts, were also reported. The CEP (2003) raised issues related to the high number of districts absorbing the costs of the policy initiative as first round funds are expended and grants are frozen or decreased. The CEP and other organisations have also noted the problems with state self-reporting. Some states struggle to establish reliable and valid testing and reporting systems, with variable levels of achievement taken as meeting the legislated performance benchmarks; and lack reliable data that would enable meaningful longitudinal comparisons of systemic interventions.

In a major study for the Harvard Civil Rights Project, Lee (2006) used the US National Assessment of Educational Progress NAEP national testing data to reanalyze states’ performance claims. Lee’s findings show no significant positive impacts on NAEP reading or mathematics achievement since the inception of NCLB, with flat or slight declines in reading achievement. Despite some positive transient improvement in Grade 4 mathematics after NCLB, these positive effects diminished and achievement returned to pre-NCLB rates after Grade 4. Lee found that there were no signs that the achievement gaps for at risk groups were diminishing. In itself this finding is troubling, but perhaps of more broad concern are the plateau effects that can be seen in the results of those states that moved toward test-based accountability systems in the 1990s (e.g. Texas, North Carolina, and Florida). Lee’s analysis concludes that state achievement tests tend to significantly inflate proficiency levels and underestimate the ongoing racial and social achievement gaps across all states: “The higher the stakes of state assessments, the greater the discrepancies between NAEP and state assessment” results (p. 11). It is unsurprising that cases of state-level test score misrepresentation and, in instances, outright fraud are currently investigation in six states, with media reports of numerous other cases.

**Phonics and the standardisation of teaching**

There are common themes between NCLB and our own recent responses to improving literacy standards and outcomes in Australia. What is distinctive about theses current debate, across both contexts, has been the rise of a ‘gold standard’ of evidence-based research as the major criterion for deciding what will be considered ‘valid’ as evidence of success in literacy teaching. The gold standard for NCLB, set within the foundation reports of the National

---

2 While there are significant challenges faced in supporting the many schools that have been identified for improvement, most districts received less federal Title 1 funding in 2004-5 than in previous years of NCLB, with further cuts for 2006-7 and 2007-8 already a reality or flagged to become so. (http://epsl.asu.edu/epru/articles/EPRU-0504-121-OWL.pdf A press release for the Centre on Education Policy released on March 23 10.00 am)
Reading Panel (2000) and affiliated studies (Snow, Burns & Griffin, 1998), is based on a number of propositions.

However, it begins from what we term the *phonics hypothesis*: that there is scientific evidence that literacy achievement can be improved through systematic curricular approaches to pedagogy that emphasise ‘alphabetics’ or phonics. In a major and award-winning reanalysis of reading research, Scott Paris (2005) claims that “most of the scientific evidence about reading skills and reading development, particularly relating to decoding skills, is based on inadequate theories, measures and interpretations” (p. 201). He also encourages a broadening of our understandings of what ‘valid’ reading assessments might be and claims that this requires new theories of reading skills and development to achieve. He argues, “alphabetics” are of importance and are necessary but not sufficient for sustained reading gains. It is important to note that once the *phonics hypothesis* is tabled, the inverse of this proposition is taken as fact: that other methods that have not been verified by the ‘gold standard’ of randomized field trials actually contribute to current patterns of early literacy achievement and failure and, specifically, the underperformance of minority and lower socio-economic groups. The Australian report released in 2005 makes precisely such a claim (National Inquiry into the teaching of Reading, 2005). While the *phonics hypothesis* is defended with examples of research which reach the ‘gold standard’, none of the studies cited in any of the reports or panels offers comparable quality evidence that other methods contribute to failure.

The next assumption of NCLB is that the ‘right method’ to improve literacy outcomes overall and minority and at risk student performance more specifically can best be achieved through what we term the *standardised curriculum hypothesis*: that standardized curriculum programs that script, monitor, and benchmark teachers' everyday teaching can be implemented across schools, communities, and student cohorts to achieve a better and more uniform spread of the optimal ‘method’ for teaching literacy.

These two assumptions—one about the efficacy of a phonics approach and another about the efficacy of standardising ‘methods’ as a way to ‘teacher-proof’ the curriculum — lead to a particular reform response and limit others. The argument looks something like this:

1. That current teacher methods for teaching literacy are unscientific and flawed (problem);
2. that government identification and selection of a scientifically verified approach to early literacy training that emphasises phonics (policy),
3. implemented through an accountability system based on state standardised testing (policy)
4. will lead to test score gains, particularly of those children from historically underperforming groups (outcome).
So the problem is seen to be teacher failure to implement a scientifically verified method. There is little recognition of the host of contributing factors identified in ethnographic, case-based, and quantitative literacy research. Factors like home/school transitions and access; the variable impacts of community cultural and linguistic background; the effects of poverty; the increasing incidence of special needs (e.g. Gregory, 2000); and the impacts of differential school resourcing (OECD, 2005) and internal tracking structures of schools (e.g. Oakes, 1985). Because this corpus of work is multidisciplinary and does not focus principally on pedagogic method as dependent variable, it was ruled out from the scientific “gold standard” of the national reports.

With teachers becoming the ‘problem’, government control becomes the solution. So it is not teacher, students, or communities who act to address the problem. Instead, the focus is upon the role of government in selecting, sanctioning and implementing the proper ‘method’ to teach and test.

**Increased accountability does not necessarily lead to more equitable outcomes**

A basic proposition of NCLB is that accountability as standardised testing is an optimal means of encouraging literacy gains especially amongst students from minority and lower-socioeconomic backgrounds. However research has suggested that this proposition may not be valid. Amrein and Berliner (2003) found that there were no consistent US state-by-state effects on the National Assessment of Educational Progress (NAEP) test results. In this large scale quantitative comparison of state results, there was no evidence that children’s literacy outcomes were improving as a result of new testing regimes of NCLB. In a reanalysis of state test scores, Rosenshine (2003) also found no consistent effects demonstrated in relation to improved outcomes or standards. While there is evidence that supports the claim that grade 8 mathematics results have improved as a result of testing (Braun, 2004; Carnoy & Loeb, 2002), there are also claims that the some results are being inflated by either exclusion of lower achieving students (Amrein & Berliner, 2003) or as a result of rising adolescent drop-out rates (Heubert & Hauser 1999).

The assumption of NCLB is that incentives and sanctions on schools and teachers based on student performance on tests will have a tonic effect on overall standards and achievement. So the assumption is that teachers and students will work harder and produce better results when faced with incentives and discipline (Nichols, Glass & Berliner, 2005). The complex factors that can be shown empirically to mediate student achievement include issues of content and construct validity and test preparation, the overreliance on single-shot assessment, and the dynamics of spatialized poverty and demographic change. Triangulated by extensive qualitative documentation on the unintended effects of NCLB (Nichols & Berliner, 2006), reanalysis of NAEP data suggests that “the relationship between high-stakes testing and its intended impact on achievement is mixed and inconclusive” (Nichols, Glass, & Berliner, 2005, p. 2).
Nichols et al. (2005) measured and ranked states according to state-level testing pressure through a system of Accountability Pressure Ratings (APR). This rating system was used to query whether “the pressure of high-stakes testing increases achievement” (p. 3). Simply, the positive link claimed by many states between the introduction of high-stakes testing through NCLB and improved student achievement is tenuous, with no gains being demonstrable on reading achievement in years 4 or 8. Like other researchers (e.g. Carnoy & Loeb, 2002) Nichols et al. (2005) did find positive relations between the APR and early math achievement of African American students. There was a negative correlation between increased accountability pressure and retention to senior and college years. That is, systemic increases in high stakes testing have an apparent link with increased drop-out rates.

**Demystifying the ‘early literacy’ solution**

NCLB is premised upon the assumption that early intervention and acquisition of “alphabetics” has longitudinal effects. In his reanalysis of the empirical research bases on early intervention, Paris (2005) makes an empirical and theoretical case that the popular push back to phonic programs as reading instruction has misconstrued the longitudinal and developmental effects of early achievement of the “constrained skills” of alphabetics. He argues that achievement of the “unconstrained skills” of comprehension—including vocabulary knowledge, inference, and critical analysis—are much stronger and robust predictors of later academic achievement. The same case is put strongly by Calfee (2003), whose analysis of California state test data noted that lower socio-economic, linguistic, cultural minority, and other at risk students experience significant problems in the transition to secondary school, even where early intervention programs have been put in place. There is a range of plausible explanations for this phenomenon, with traditional comprehension research suggesting the key role in vocabulary knowledge in sustainable upper primary and middle years reading achievement (for an overview, see Alvermann, 2002). There is, further, a general consensus in the literature on middle years of schooling that at risk students would benefit from systematic and ongoing instructional foci and school diagnostic and remediation resources in literacy (for a review, see Luke, Elkins et al., 2003).

**Lessons from US Policy**

We conclude with several lessons from US policy.

* First, the combination of increased testing, standardised programs, increased accountability and incentives/sanctions for schools, districts and states who do not reach targets has not been a success. As we have shown here, while states reported test score gains for at risk students – these gains were not confirmed in reliable, national sample testing.

---

3 However the same positive relationship was not evident in early reading within the same cohort.
Second, extensive local ‘collateral effects’ have been documented (e.g. Nichols et al, 2005). These include: test-preparation sessions, school and district-level test administration and test-score manipulation, loss of experienced teachers, inadequate funding for professional development, lack of support for English language learners and students with special needs, increased teacher utilisation of packaged materials without ‘scientific’ backing, and, significantly, a narrowing of the overall curriculum.

In a recent address to the European Reading Conference in Berlin, Andreas Schleicher (2007) of the OECD argued that Australia’s PISA literacy results placed it in the top tier of countries – and that this was the consequence of high levels of professionalism. There are indeed important and unresolved issues and challenges in Australian and Queensland literacy education. And we have made major, internationally recognised progress since the initiation of policies like Literate Futures (2001). Ontario, New York and other jurisdictions have adopted the “four resources” model for balanced reading programs.

The progress we have made has been the result of policies that set the conditions for increased teacher professionalism with accountability – an element that Schleicher views as crucial to the success of countries like Finland, Canada and Australia. What is clear is that the US agenda – with high levels of centralised prescription of curriculum and pedagogy, and increased high stakes testing – is not working.

References


Centre on Education Policy (2003) *From the Capital to the Classroom: Year 3 of the No Child Left Behind Act*. Washington, DC: Centre for Educational Policy.


