

Chapter 2

Large-Scale Test Data: Making the Invisible Visible

Gilah Leder and Sarah Lubienski

Introduction

Information on pupil performance is key to the successful implementation of targeted education policies and it is not surprising that in the past two decades national tests have emerged as an important tool for providing a measure of educational achievement. (Figel 2009, p. 3)

Formal assessment of achievement has a long history. Official written examinations for selecting civil servants were already in use in China more than 3000 years ago (Kenney and Schloemer 2001). Educational assessment is a far more recent practice, with its genesis commonly traced to the nineteenth century. Over time, the development of large-scale, high-stakes testing and explorations of its results have proliferated. “Many nations,” wrote Postlethwaite and Kellaghan (2009), “have now established national assessment mechanisms with the aim of monitoring and evaluating the quality of their education systems across several time points” (p. 9). In some countries, the practice is limited to a number of core curriculum subjects but in others the testing regime is broad.

In this chapter, we confine our attention to large-scale tests used to measure mathematical progress and proficiency in two countries: Australia and the USA. We examine the aims, capacity, and limitations of the tests, what they can tell us about student performance and what can be learnt from an international comparison. In the Australian setting we focus primarily on the National Assessment Program—Literacy and Numeracy (NAPLAN). For the American context we have chosen to focus initially on the National Assessment of Educational Progress (NAEP). To supplement the findings from these tests limited reference is also made to three other tests: the (American) Early Childhood Longitudinal Study (ECLS), the Trends

G. Leder (✉)

Faculty of Education, Monash University, Clayton, Australia
e-mail: gilah.leder@monash.edu

S. Lubienski

Department of Curriculum and Instruction, University of Illinois, Urbana-Champaign, USA
e-mail: stl@illinois.edu

in International Mathematics and Science Study (TIMSS), and the Programme for International Student Assessment (PISA).

How justified is our faith in using large-scale test results as a catalyst for change, for dealing constructively with diversity, and for enhancing social inclusion?

Early Developments

The USA

In a submission to America's National Assessment Governing Board (NAGB), Vinovskis (1998) sketched the motivations of early American advocates for the collection of comparative educational data as follows:

Nineteenth-century reformers had an abiding faith that the compilation and display of numerical data not only would reveal the inherent regularities in behavior, but also would suggest possible options for making changes. They believed that if policymakers and the public were presented with the appropriate comparative data on social reforms such as education, they would soon want to improve their own policies accordingly. (Vinovskis 1998, p. 3)

These sentiments, we show in this chapter, still underpin—within America, Australia, and indeed more widely—contemporary preoccupations with large-scale testing and its putative benefits.

Much debate, political manoeuvring, and balancing of competing concerns and interests of local, state, and federal bodies preceded the creation and eventual introduction in the USA of the NAEP tests in the mid-1960s (see, e.g., Vinovskis 1998; Jones and Olkin 2004). The tests are now administered regularly to representative samples of students at grades 4, 8, and 12.

Australia

For many years, Australian states and territories ran their own numeracy and literacy testing programs. Although much overlap could be found in the assessment instruments used in the different states, there were also variations—some subtle, others substantial—in the tests administered. Finally, in 2008, a serious program of national testing was launched. The NAPLAN tests represented a significant turning point in Australia's educational system. For the first time, students in years 3, 5, 7, and 9, irrespective of their geographic location in Australia, sat for a common set of tests.

More about NAEP and NAPLAN

NAEP

Every 2–4 years, NAEP assesses the mathematics knowledge and attitudes of large, representative samples of US students at grades 4, 8, and 12 (roughly ages 9, 13,

and 17, respectively). The Main NAEP mathematics results are reported via overall scale scores and proficiency levels (basic, proficient, advanced). Scale scores are also available for each of five mathematical strands: (1) number/operations; (2) measurement; (3) geometry/spatial sense; (4) data analysis, statistics, and probability; and (5) algebra/functions. Additionally, NAEP administers student and teacher surveys to collect information about students' mathematics attitudes and experiences, as well as teachers' backgrounds and classroom practices.

Several features complicate the analysis of NAEP data. The assessment uses multi-staged, stratified random sampling (geographic areas, then schools, and then students are selected). Since NAEP is designed to provide a snapshot of national achievement as opposed to providing feedback to individual students or schools, each student is administered only a subset of the NAEP items. This allows NAEP to monitor national performance on a rich variety of items without over-burdening individual students. These complications are addressed with several techniques, including the use of student sampling weights and imputed achievement values. For further information about the structure of NAEP data, see Johnson (1992) or Johnson and Rust (1992).

Reporting NAEP Data

The National Center of Education Statistics publishes general NAEP results for the USA and for key demographic subgroups. Scores are reported for the nation and each state—scores for individual students, teachers, and schools are not available. State tests—not NAEP—are typically used to track the progress of individual US students and schools.

NAEP scores for grades 4 and 8 are on a common (cross-grade) scale of 0–500. Since 2005, grade 12 scores are on a 0–300 scale. Most of the NAEP analyses discussed here were conducted via NAEP's web-based data tool available at <http://nces.ed.gov/nationsreportcard/naepdata/>. This tool allows users to examine mathematics achievement and survey data, and to make comparisons by demographic variables, including socioeconomic status (SES), gender, race/ethnicity, and home language. Raw NAEP data are also available to researchers via an application process.

NAPLAN

The NAPLAN numeracy tests contain multiple choice and supply response items. Their scope and content are informed by the Statements of Learning for Mathematics (Curriculum Corporation 2006) and cover four broad, and sometimes overlapping, numeracy strands: algebra, function, and pattern; measurement, chance, and data; number; and space. Since students have 40 or 50 minutes (depending on grade level) to complete a numeracy test paper, the content and scope of these high-stakes tests are inevitably limited.

Reporting the Data

Students' NAPLAN numeracy scores for years 3, 5, 7, and 9 are reported on a common scale, based on the Rasch model (see, e.g., Andrich 1988), so that any given scale score represents the same level of achievement over time. The use of a common scale for all domains measured enables changes in individual student achievement to be tracked, and provides a longitudinal dimension to the data. For each year level, the proportion of students with scores in the six proficiency bands deemed appropriate for that level is provided. For year 3, the bands are 1–6; for year 5, 3–8; for year 7, 4–9; and for year 9 bands 5–10.

Each year, summative results of the NAPLAN tests are published in some detail and are made available to the public. Particular attention is paid to the proportion of students meeting, or failing to meet, the specified minimum standards, and to mean NAPLAN scale scores. Individual student data are released to the relevant school attended, and the student's parents; unless special consent has been obtained, they are otherwise unavailable.

As well as the aggregated data, results are reported separately by: state/territory, gender, Indigenous status, language background status (language background other than English; LBOTE and non-LBOTE), geolocation (metropolitan, provincial, remote, and very remote), parental educational background, and parental occupation. These factors overlap with those frequently used as descriptors of equity/inequity. The categories are not mutually independent and can have a simple or compounding impact on students' NAPLAN scores. At the same time, as noted in Chapter 6, such data can, indirectly and subtly, reinforce prevalent stereotypes. More generally, as we show in this chapter, the publication of results in this way can highlight advantages and disadvantages linked to situational and external factors.

Beyond NAEP and NAPLAN

Australia and the USA also participate in two highly influential large-scale international tests of mathematics achievement: the TIMSS and the PISA. In both countries, students' results on the national and international tests attract much attention, within and beyond the educational community (see, e.g., Carnoy and Rothstein 2013; Thomson 2010; Thomson et al. 2011). In the most recent TIMSS tests (see Mullis, Martin, Foy and Arora 2012), the average scale score for Australian students at the fourth grade and eighth levels in TIMSS was somewhat below that of students in the USA—25 points at the former and 4 points at the latter grade level. On the PISA 2012 test, the mean score of Australian students (504) was above the international mean score; that of American students (481) was below the international average.

Because of differences in the scope of the tests and in the student groups to whom the large-scale tests are administered, complementary as well as overlapping data are yielded by the various testing instruments: NAEP, NAPLAN, TIMSS, and PISA.

Comparing the Design of NAPLAN, NAEP, TIMSS, and PISA

NAPLAN

As described above, the NAPLAN testing regime is aimed at students in years¹ 3, 5, 7, and 9. Participation in the tests is formally voluntary. Compliance, however, is high. For example, in 2011 approximately 95% of the Australian year 3 cohort and 92% of the year 9 cohort completed the NAPLAN numeracy tests. Thus, NAPLAN is best described as a census test. The NAPLAN numeracy tests provide information about student performance, both individually and for specific groups, on traditional mathematical tasks taught in class. There is no attempt to measure students' attitudes or to probe student knowledge not readily measured within the short time slot allocated to the test.

NAEP

NAEP is administered regularly to selected students in grades 4, 8, and 12, with the sample chosen to be nationally representative with respect to factors such as ethnicity, economic background, geographic location, and school size. Only a small percentage (less than 5%) of US students in those grades participate. Still, the sample sizes are large, with roughly 175,000 at grades 4 and 8 and 50,000 at grade 12 participating in recent assessments.² Almost 100% of selected fourth and eighth graders at public schools take the test, with participation rates lower for twelfth graders (83%) and private school students (70–75%).³ The mathematical content assessed by the “Main NAEP” (discussed thus far and throughout the chapter) reflects current thinking about the most important curricular topics and includes a mix of multiple choice, short answer, and extended response items.⁴

TIMSS and PISA

The TIMSS tests, aimed at students in years 4 and 8, and the PISA tests administered to 15-year-old students, are restricted in Australia to “a light sample (of) about

¹ Terminology (years or grades) as used in the tests.

² Grade 4 and 8 samples are particularly large because they are selected to be representative of each US state and then aggregated to be nationally representative. The grade 12 sample, in contrast is simply nationally representative and therefore smaller.

³ These and other details are available at: http://nationsreportcard.gov/math_2009/about_math.asp, http://nationsreportcard.gov/math_2011/about_math.asp, and http://nces.ed.gov/nationsreportcard/tdw/sample_design/.

⁴ The “Long-Term Trend” NAEP (administered periodically with far smaller sample sizes) is a more traditional, multiple-choice test that tracks US students' mathematics and reading knowledge on the content considered important when it was begun in the early 1970s. We focus in this chapter on the more widely referenced and discussed “Main NAPLAN” data.

5 % of all Australian students at each year or age level” (Thomson 2010, p. 76). A similar sampling approach is used in the USA, with roughly 5000–8000 students (less than 1 %) in each cohort assessed (<http://nces.ed.gov/timss/faq.asp>; Carnoy and Rothstein 2013).

The aim of the TIMSS testing program is ambitious. Its scope is typically discussed in terms of three levels of the curriculum. These are:

the intended curriculum (what society expects students to learn and how the system should be organised to facilitate this), the implemented curriculum (what is actually taught in classrooms, who teaches it and how it is taught) and the achieved curriculum (which is what the students have learned, and what they think about these subjects). (Thomson 2010, p. 76)

Thus, the TIMSS tests assess how well students can handle the work taught in class, as well as a range of broader situational and background factors.

The PISA tests have a different focus. The mathematical component of this test aims to assess not only how well students have mastered mathematical content but also how well they can apply that knowledge to real world settings.

National Tests: Intended Benefits

Australia

The reputed benefits of a national testing program were widely discussed and disseminated prior to the introduction of the NAPLAN tests. They continue to be repeated in official documents published by the Australian Curriculum Assessment Reporting and Curriculum Authority (ACARA). Advantages of the test agenda which continue to be lauded by its supporters mirror those commonly put forward in the wider literature: assessment consistency across different constituencies, increased accountability, and a general driver for improvement.

The national tests, which replaced a raft of tests administered by Australian states and territories, improved the comparability of students’ results across states and territories...

Australians can expect education resources to be allocated in ways that ensure that all students achieve worthwhile learning during their time at school...

All Australian schools benefit from the outcomes of national testing, with aggregated results made available through comprehensive reports at the national and school level, accessible on-line. (excerpts retrieved from ACARA 2011a)

USA

Similar to the goals of NAPLAN, NAEP is commonly referred to as “the nation’s report card.” Unlike the variety of state and district tests that continue to be administered in US schools, NAEP is the only ongoing assessment of students representing the US. Given NAEP’s inclusion of demographic and instruction-related variables, NAEP is also designed to monitor US students’ learning experiences and disparities in academic outcomes.

National Tests: Critics' Concerns

There are, inevitably, those who question the benefits of large and high-stakes testing programs and express unease about their impact. Their voices, too, must be acknowledged. Criticisms often expressed

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 et al. 2012, p. 4)

The tendency, in both Australia and the USA, to use published test data to make glib and indefensible comparisons between schools has also caused dismay. The push for teacher and administrator accountability based on standardized test results has attracted much condemnation and has led to recent, high-profile cases of “cheating” among the staff in some US and Australian schools (Bachelard 2011; Strauss 2013; Winerip 2013).

Additionally, some scholars argue that standardized tests are inappropriate measures of achievement that often hinder instead of help efforts toward equity. For example, Gutiérrez (2008) has argued that standardized tests are overly narrow measures and that repeatedly calling attention to achievement gaps between groups while ignoring within-group variation only serves to confirm stereotypes.

What *can* be Learned from the NAPLAN and NAEP Tests?

Our emphasis in this section is on productive ways of interpreting the data, of using the published results to raise questions and issues which warrant further investigation, and focusing on current inequities that cry out for positive interventions. We begin by listing a number of indicative examples taken from the NAEP and NAPLAN tests, and discuss what can, and cannot, be inferred from the results. Where useful, we refer to data beyond these tests.

In this way we not only examine the claims that large-scale tests contribute to increased accountability in the educational sectors and can serve as general drivers for improvement, but also highlight issues which merit reexamination and further explorations.

We start with NAEP and NAPLAN test data to examine the following questions:

- Has student mathematics achievement changed over time? Are any patterns observed unique to mathematics?
- Do we obtain any additional information if the test data are reported separately for different groups? In this chapter, we consider test outcomes by race/ethnicity (NAEP data only), language background, indigeneity (NAPLAN data only), gender, and geolocation. The impact of geolocation on students' achievement, in and beyond NAPLAN, is also addressed, and in some detail, in Chapter 3 of this volume.

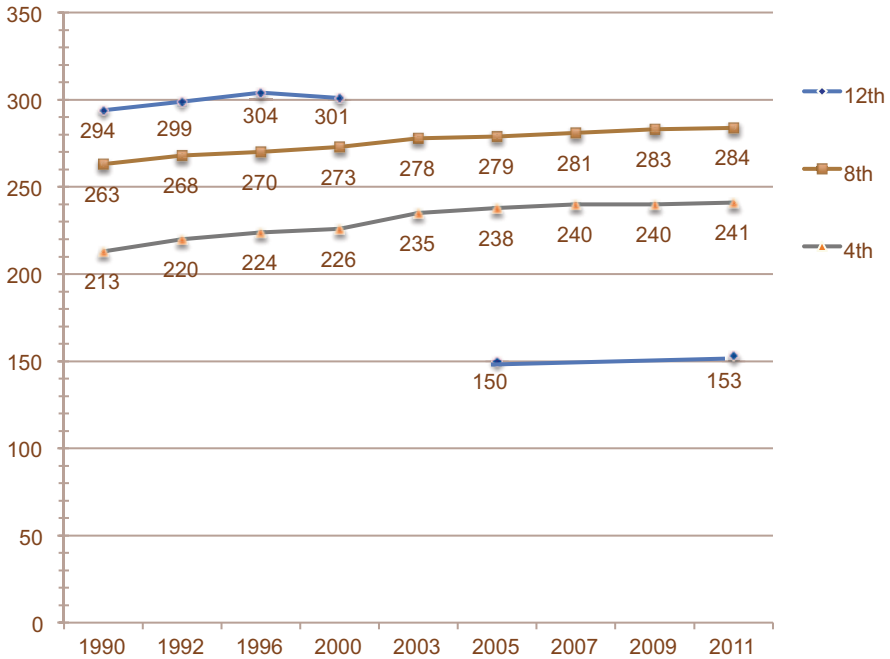


Fig. 2.1 Main NAEP mathematics results, 1990–2011, grades 4, 8, and 12

- What can we learn from item-level analyses?
- How can national survey data on student affect and instruction help us understand disparities between groups?

Achievement in Mathematics Over Time

Main NAEP Results 1990–2011

Since 1990, when the current Main NAEP Framework was established for grades 4 and 8, achievement has substantially increased at those grades. The standard deviation varies by year and grade level, but averages about 30 points at grade 4 and 36 points at grade 8. As shown in Fig. 2.1, increases are large, consisting of roughly 0.6–0.9 standard deviations. Gains at grade 12 are less evident, although the change in framework and reporting scale makes it difficult to draw conclusions about trends at that grade level. In contrast, during this same time period, reading scores increased only about 0.1–0.2 standard deviations at grades 4 and 8.

One of the most plausible explanations for the upward trend in mathematics is that the NCTM Standards movement began in 1989, and the framework for the 1990 NAEP assessment was designed to be aligned with those standards. That

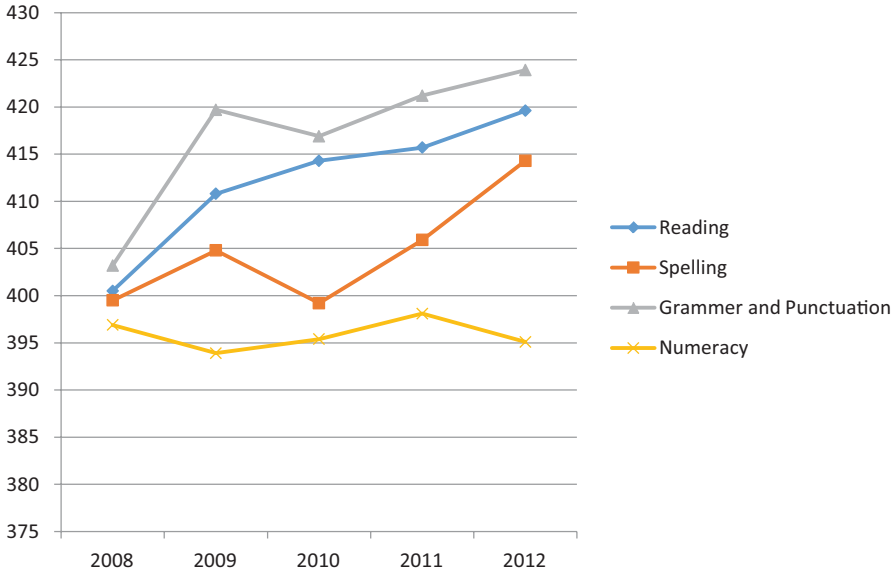


Fig. 2.2 NAPLAN Numeracy scores compared to three NAPLAN literacy measures

framework included topics that had typically not been included in the elementary and middle-school curriculum, such as probability, data analysis, and algebra. The NAEP framework also emphasized mathematical reasoning and communication more than prior assessment frameworks. The reforms influenced state and district standards and assessments (Usiskin 1993), as well as curriculum materials and teacher education programs. US students began having the opportunity to learn the content emphasized on NAEP, and this can be seen in the score gains during the 1990s and perhaps beyond. Another theory to explain gains made in the beginning of 2003 is that the new federal law, “No Child Left Behind,” began to shift toward more high-stakes uses of NAEP and other standardized tests, which put pressure on educators to prepare students for such tests and to pay particular attention to the performance of traditionally underserved student groups.

Main NAPLAN Results for 2008–2012

For NAPLAN, the time span for monitoring test results is—at the time of writing this chapter—limited to 5 years.

At each of the year levels tested, only one numeracy test is administered, compared with four different tests for literacy: reading, writing, spelling, and grammar and punctuation. The mean test scores for four of the five national achievement scales at the year 3 level are shown in Fig. 2.2. The data for writing have been omitted because of a change to this test over the life of the NAPLAN testing regime,

making it inappropriate to report longitudinal data from this test. Initially, for the period 2008–2010, the writing results were reported on the narrative writing scale; in 2011–2012 this was changed to the persuasive writing scale to reflect more effectively the range of what is required in the curriculum.

From Fig. 2.2, it can be seen that numeracy scores have remained steady, while literacy scores have generally increased over the 2008–2012 period. Space restrictions prevent detailed presentation of the year 5, 7, and 9 data. What can be said, however, is that the pattern evident at the year 3 level of relatively steady numeracy scores and somewhat increased scores between 2008 and 2012 for the three literacy measures is largely repeated at the other 3 year levels.⁵ This is an intriguing finding. Are the different numeracy and literacy findings a function of the nature of the different tests and the way they capture curriculum content, or does the explanation lie elsewhere? Why, it should also be asked, do the achievement patterns for grade 4 students on the NAEP tests and year 3 students on the NAPLAN tests differ?

Achievement in Mathematics of Different Student Groups

NAEP Achievement by Race/Ethnicity and Sex

The impressive gains in NAEP scores since 1990 raise the question of whether all US students made similar gains, or if gaps between more and less advantaged students increased or decreased during this time period.

NAEP offers information on many different subgroups. The three largest racial/ethnic subgroups for which NAEP reports data are White, Black, and Hispanic students.⁶ Figure 2.3 below reveals that White, Hispanic, and Black students all made fairly similar gains between 1990 and 2011. However, the steepest gains for Black and Hispanic students were made between 2000 and 2003, leading many politicians to credit the “No Child Left Behind” reforms for the narrowing of gaps. The patterns at grade 8 were similar.

There are additional patterns to note in Fig. 2.3. First, throughout the years, White students consistently scored 0.6–1.0 standard deviation higher than Hispanic and Black students. These are large, persistent disparities that merit continued attention. However, these same data show that such disparities are not inevitable. In fact, it took less than two decades for Black and Hispanic students to surpass a mean score of 220, which was the mean for White students in 1990. In other words, if White students had maintained their initial 1990 scores while mathematics learn-

⁵ See, e.g., 2012 NAPLAN National Report, retrieved from www.nap.edu.au/verve/_resources/NAPLAN_2012_National_Report.pdf.

⁶ Although there are differences of opinions about appropriate categories and terms to use to describe various groups in the USA, we use NAEP’s school-reported categories and terms when describing NAEP results for racial and ethnic subgroups. Latino/a students are included as “Hispanic” and are generally not included in the “White” and “Black” categories. NAEP also includes a category for “2 or more races,” but we do not report on that relatively small category here.

Diversity in Mathematics Education

Towards Inclusive Practices

Bishop, A.; Tan, H.; Barkatsas, T.N. (Eds.)

2015, XVIII, 262 p. 32 illus., 15 illus. in color., Hardcover

ISBN: 978-3-319-05977-8