LONG-RUN TRENDS IN SCHOOL PRODUCTIVITY: EVIDENCE FROM AUSTRALIA

Andrew Leigh
Research School of Economics
Australian National University
ACT 0200, Australia
andrew.leigh@anu.edu.au

Chris Ryan
(corresponding author)
Research School of Economics
Australian National University
ACT 0200, Australia
chris.ryan@anu.edu.au

Abstract

Outside the United States, very little is known about long-run trends in school productivity. We present new evidence using two data series from Australia, where comparable tests are available back to the 1960s. For young teenagers (aged 13–14), we find a small but statistically significant fall in numeracy over the period 1964–2003 and in both literacy and numeracy over the period 1975–98. The decline is in the order of one-tenth to one-fifth of a standard deviation. Adjusting this decline for changes in student demographics does not affect this conclusion; if anything, the decline appears to be more acute. The available evidence also suggests that any changes in student attitudes, school violence, and television viewing are unlikely to have had a major impact on test scores. Real per child school expenditure increased substantially over this period, implying a fall in school productivity. Although we cannot account for all the phenomena that might have affected school productivity, we identify a number of plausible explanations.

1. INTRODUCTION

All too frequently, education policy debates focus on inputs rather than outputs. To a large extent this is a function of the available data. While input measures such as class size, teacher salary, and per student funding are readily comparable over time, output measures such as test scores are often not designed to be compared across years (either because they are re-standardized each time they are administered or because the test instrument itself is changed from year to year).

Yet from a policy-making perspective, what should matter most is school productivity—the results a school system achieves for a given level of inputs. As with other productivity measures in the economy, it is reasonable to expect that these should steadily improve over time, as new innovations allow schools to produce better results with the same level of real inputs.

In this article, we analyze long-run trends in school productivity in Australia, comparing standardized test score results in literacy and numeracy with data on per student expenditure. Combining data from two nationally representative sets of tests, we are able to compare numeracy scores from 1964 to 2003 and literacy scores from 1975 to 1998.

Although literacy and numeracy scores are by no means the sole output of the education process, equipping children with good reading, writing, and mathematical skills is nonetheless an important function of schools. At an individual level, studies have shown that Australian students with better literacy and numeracy in their early teenage years are more likely to be employed in their twenties (Marks and Fleming 1998a) and that, conditional on being employed, they tend to earn higher wages (Marks and Fleming 1998b). At a national level, countries where the labor force is more numerate tend to grow faster than countries with lower levels of numeracy (Hanushek and Kimko 2000; Jamison, Jamison, and Hanushek 2006).

Only a small number of studies has looked at changes in test scores and spending over several decades. For the United States, Hanushek (1997) showed that test scores were essentially flat over the period 1970–94, while real per child expenditure grew by 2.5 to 3 percent per year. Hoxby (2002) demonstrated that this finding held true for the period 1970–98 and that adjusting for student demographics made little difference to the overall pattern. Similarly, Gundlach, Woessmann, and Gmelin (2001) find that test scores in eleven developed countries were basically flat over the period 1970–94, while real per child funding rose dramatically. However, because their findings rely on benchmarking against the U.S. National Assessment of Educational Progress (NAEP), these results are partly driven by trends in U.S. NAEP scores.

Our analysis of school productivity follows an extensive literature on public sector productivity (see, e.g., Rosen 1993; Atkinson 2005; Boyle 2006; Douglas

2006; Weale 2007). As Atkinson (2005, p. 18) noted, "There is a strong case for devoting significant resources at this time to improving the measurement of public sector output on account of its increased saliency in policymaking and public debate." However, as Rosen (1993) has pointed out, one hesitation that arises when assessing public sector productivity is choosing the appropriate measure. Since no metric perfectly captures all aspects of the effectiveness of a public sector organization, it is important to recognize that any chosen output measure is merely a proxy variable. Notwithstanding their limitations, studies that have analyzed educational productivity have typically opted to use test scores as their primary output measure (see, e.g., Atkinson 2005). For example, the United Kingdom Department for Education and Skills (2005) noted that it had considered other output measures, such as school inspections, but concluded that test scores were preferable on the basis of data availability, transparency, and their potential to be linked to labor market outcomes. The Australian Productivity Commission (SCRGSP 2007) lists its preferred outcome measures as test scores for reading, writing, numeracy, science, and civics, plus technological literacy, vocational education in schools, completion, destination, and "other areas to be identified."[1]

As far as we are aware, ours is the first study to look at long-run trends in literacy and numeracy without relying on the U.S. NAEP. For U.S. policy makers, Australia is a useful laboratory in which to consider changes in school performance because while its labor market has much in common with the U.S. labor market, its school system has a number of features that distinguish it from the United States. Chief among these are that schools are funded by the federal and state governments (and not by local property taxes), around one-third of pupils attend nongovernment schools, and decisions about teacher appointment and remuneration in government schools are made at a state level.

To preview our results, we find no evidence that test scores have risen over the past four decades, and some evidence that scores have fallen. This finding is consistent with earlier studies using some of the same data sets (Afrassa and Keeves 1999; Rothman 2002). We explain how this finding can be reconciled with evidence on the Flynn effect (growth in measured intelligence quotient [IQ] scores over the twentieth century in developed countries) and with Australia's ranking on international tests. We also find that adjusting

1. Our analysis does not use school completion rates as a metric of school quality, on the basis that students' decisions to obtain more schooling do not necessarily reflect higher school productivity. In the Australian context, youth labor market conditions have been shown to be extremely important. Ryan and Watson (2004) noted that the parameters of the regression equation they estimated, in conjunction with the change in variable values, would have attributed the entire increase in Australian year 12 retention rates from the 1970s to the 1990s to the decline in full-time job opportunities for teenagers.

for the observed demographic characteristics of students does not affect this conclusion; if anything, the decline appears to be more acute. We also review other possible explanations, such as changes in early school-leaving behavior, and conclude that they are unlikely to have significantly affected trends in test scores.

Turning to education expenditure, we observe a significant increase in per child spending from the mid-1960s to the early 2000s. If we measure school productivity in terms of the average test score divided by the average real per child expenditure, this implies that school productivity has fallen over the past four decades. Using our preferred measure of school input prices, we estimate that school productivity has declined by 12–13 percent between 1975 and 1998 and by 73 percent between 1964 and 2003.

The remainder of this article is structured as follows. Section 2 outlines the test scores that we use and presents the basic trends in literacy and numeracy. Section 3 examines the extent to which the observed patterns might be due to changes in demographics or other factors beyond the school gate. Section 4 looks at how per child spending has changed. The final section concludes.

2. TRENDS IN LITERACY AND NUMERACY

Longitudinal Surveys of Australian Youth

The first set of tests that we use to analyze changes in literacy and numeracy are data from four Longitudinal Surveys of Australian Youth (LSAY) cohorts. These cohorts are the Youth in Transition 1961 and 1975 birth cohorts (YIT 61 and YIT 75) and the LSAY 1995 and 1998 grade 9 cohorts (LSAY 95 and LSAY 98). Those in the first cohort took the literacy and numeracy tests in 1975, while those in the last cohort took the tests in 1998. Although subsequent cohorts of the LSAY have been surveyed, the test instrument is not comparable to that used in earlier cohorts, and changes to the survey design mean that the subjects in the later cohorts are older.

Students in the first two LSAY cohorts took standardized tests at age fourteen, while students in the last two cohorts took standardized tests in grade 9. There are four ways to compare achievement across time using these data: using the full sample in all years, restricting the sample to fourteen-year-olds, restricting the sample to grade 9 students, or restricting the sample to those who were aged fourteen and in grade 9.

We adopt the last approach. That is, we opt to compare cohorts by focusing only on the common group: those who were aged fourteen and in grade 9. This is designed to limit the impact of changes in the age-grade composition of the surveys over time, since both students' ages and grades may affect their performance on achievement tests.

This is a different approach from the preferred specification used by Rothman (2002), who argues that using those aged fourteen and in grade 9 suffers from the problem that underperforming students were more commonly required to repeat grades in the 1970s than in the 1990s. Rothman cites Australian Bureau of Statistics (ABS) data showing that the share of fourteen-year-olds who were enrolled in grade 8 or below fell from 24 percent in 1975 to 16 percent in 1995–2000, and he argues that the average quality of the pool of students aged fourteen and in grade 9 might have been lower in the later cohorts. Therefore average school achievement for students aged fourteen and in grade 9 in the later cohorts would be expected to be lower than in earlier cohorts.

Our view is that any change in grade repetition is unlikely to have had a significant impact on trends in the LSAY. First, there were offsetting trends on age-grade distributions across states and territories. Grade repetition rates certainly fell over the period in states and territories where they were high initially. But in other states and territories there were trends toward later school commencement, especially among those who would have been youngest in their grade cohorts, and once more this phenomenon might be expected to have affected students who were of below-average ability. Thus in some states and territories, the proportion of students "old" for their grade, given the school commencement rules operating in their jurisdiction, actually increased in the later LSAY cohorts, presumably raising the average ability of fourteen-year-olds in grade 9. There were also some states and territories in which these patterns barely changed at all.

Second, empirically, the shares of the cohorts aged fourteen and in grade 9 did not change as much in the LSAY data as it did in the ABS data. The (weighted) proportion of the sample aged fourteen and in grade 9 was 60.3 percent in the YIT 61 cohort and 63.7 percent in the LSAY 98 cohort. In unweighted data, the change was just 1.5 percentage points. The reason the change in the proportions is smaller in the LSAY data is one of coverage—students' integer ages in the LSAY were recorded as of 1 October in the survey year, and in the YIT surveys only those aged fourteen as of that date were included. In the ABS data, ages are classified as of 1 July. Because of the complex pattern in which students born at different times of the year are distributed across school grades, which also differs substantially by jurisdiction, this different coverage means that the LSAY data have been much less affected by the trend identified by Rothman (2002) (and by compositional effects driven by differential population growth between states) than the ABS data.

Finally, it is possible to estimate how much such phenomena might have affected our results. First, we can identify students aged fourteen and in grade 9 who should come from similar parts of the ability distribution in the different

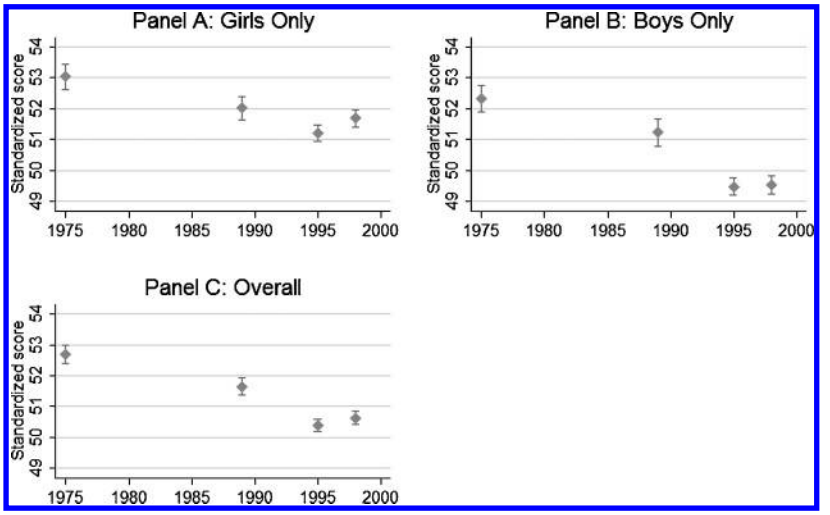

Figure 1. Literacy (LSAY). Notes: Sample is 14-year-olds in grade 9. Capped spikes depict 95% confidence interval for mean.

cohorts. These are individuals born in the same months of the year from the same state, but in different cohorts, where the proportion of the sample aged fourteen and in grade 9 was the same. It is also possible to identify cohorts of students where the distribution changed substantially in the way identified by Rothman (2002)—that is, where the proportion aged fourteen and in grade 9 increased. We find the trends we present below for the entire group aged fourteen and in grade 9 are also borne out for these other groups, but the decline in literacy and numeracy performance is most pronounced in the group where the proportion aged fourteen and in grade 9 increased the most—which is precisely the effect identified by Rothman (2002).[2]

The LSAY literacy and numeracy scales used here are taken from Rothman (2002). They were developed using the common items asked of the cohorts using a Rasch item response model. The scales were constructed to have a mean of 50 and a standard deviation of 10 across the cohorts.

Figures 1 and 2 contain trends in test scores from the LSAY, with dots representing the mean and error bars representing the 95 percent confidence interval for the mean. Across all tests, the mean of the standardized test scores is set at 50 and the standard deviation at 10. The results show that over the

2. Moreover, the trends are identical where we also exclude individuals who appear "old" or "young" for their grade given the typical age-grade level of other students in their state born in the same months of the year.

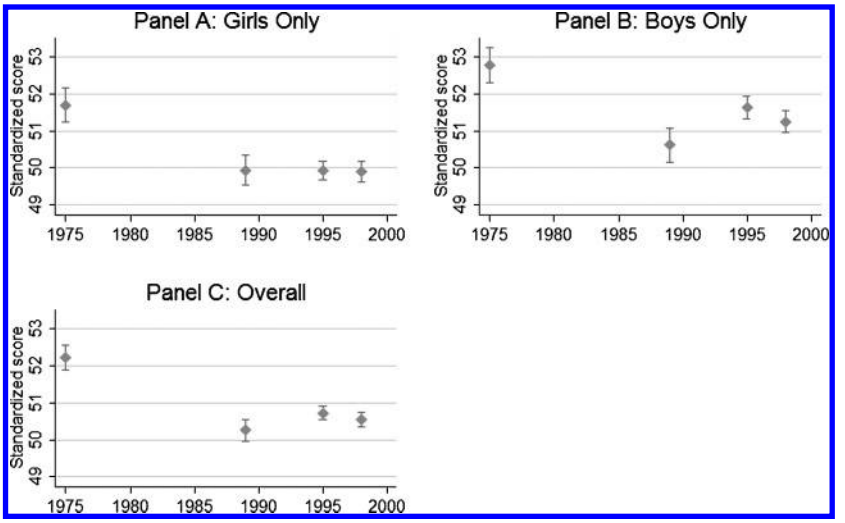

Figure 2. Numeracy (LSAY). Notes: Sample is 14-year-olds in grade 9. Capped spikes depict 95% confidence interval for mean.

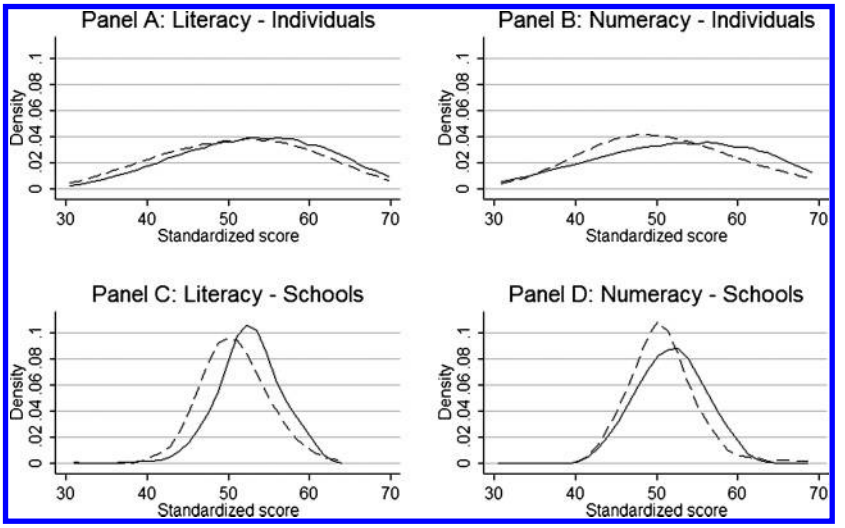

Figure 3. Distributions (LSAY). Notes: Sample is 14-year-olds in grade 9. Continuous line represents 1975; dashed line represents 1998.

period 1975–98 there was a statistically significant decline in the literacy and numeracy test scores of both boys and girls.[3]

Was the fall in mean scores driven mostly by changes in the left tail or right tail of the distribution? And how did the distribution of test scores across

3. This finding is sensitive to our choice of students aged fourteen in year 9. Results presented in Rothman (2002) show that when the comparison uses the full sample, or just fourteen-year-olds, literacy and numeracy scores are flat from 1975 to 1998.

schools change over time? To answer these questions, figure 3 presents kernel density plots (effectively smoothed histograms) of the distributions in 1975 and 1998. In these graphs, the horizontal axis shows the standardized score, while the vertical axis shows the share of observations that are at or close to each score. Panels A and B show the distribution of individual literacy and numeracy scores. In the case of individual literacy scores, it appears that the distribution has simply moved toward the left. By contrast, the distribution of individual numeracy scores has also changed shape slightly, with fewer children scoring around sixty and more scoring just below fifty.

Panels C and D show the distribution of average literacy and numeracy scores at the school level. In both cases, the graphs overlap on the far right tail, indicating a similar number of high-performing schools in both years. Elsewhere, the graphs appear to have essentially shifted to the left. For literacy, the distribution of school performance has become slightly more dispersed over time, while for numeracy, the distribution of performance across schools has become slightly more compressed over time.

IEA Surveys

The second source of data that we analyze are the five mathematics surveys conducted by the International Association for the Evaluation of Educational Achievement (IEA). These are the 1964 First IEA Mathematics Study, the 1978 Second IEA Mathematics Study, the 1994 Third IEA Mathematics Study (also known as the 1995 Third International Mathematics and Science Study, or TIMSS, though it was conducted in Australia in 1994), the 1999 TIMSS, and the 2003 TIMSS.

In comparing the cohorts, we need to take into account shifts in the composition of test takers. There are three aspects to this. First, the geographic coverage of the test steadily increased over time, with the 1964 survey covering five states (New South Wales [NSW], Victoria [VIC], Queensland [QLD], Western Australia [WA], and Tasmania [TAS]), the 1978 survey covering these plus the Australian Capital Territory (ACT) and South Australia (SA), and the 1994 and subsequent surveys also covering the Northern Territory (NT).[4] Second, the surveys extended their coverage across school types, with the 1964 survey covering only government schools and later surveys covering both government and nongovernment schools. Third, the surveys changed their age-grade coverage, with the 1964 survey covering grade 8, the 1978 survey covering thirteen-year-olds, the 1994 survey covering grades 7–9, the 1999 survey covering grades 8–9, and the 2003 survey covering grade 8.

4. Although the ACT was part of the NSW schooling system in 1964, ACT schools do not appear to have been sampled in the 1964 survey (Afrassa and Keeves 1999).

To take account of these various shifts, we use a Rasch item response model to estimate a standardized score, using common questions across the tests.[5] For ease of comparison, we then express the tests on the same scale as the LSAY results (with a mean of 50 and a standard deviation of 10). Where T is the test score of student i, tested in year y, in grade g, and of age a, we estimate the following regression:

$$T_{iyga} = \alpha + \beta_y I_y^{\text{test year}} + \gamma_g I_g^{\text{student grade}} + \delta_a I_a^{\text{student age}} + \varepsilon_{iyga}. \quad (1)$$

The parameter β_y denotes how mean scores have changed, holding constant changes in the age and grade composition of students taking the test.[6] Since the version of the 1994–2003 data sets we are using does not contain information on state or on whether the student attended a nongovernment or government school, we first estimate the model just for 1964 and 1978, with the sample restricted to the student population covered by the 1964 test. We observe a decline of 2.5 points, which is significant at the 1 percent level. We then estimate the model using data for all tests and find that the drop from 1964 to 1978 is 2.0 points (still significant at the 1 percent level). Given the similarity between these two estimates, we opt to compare across all cohorts.

Figure 4 charts the results. Over the 39 years from 1964 to 2003, we observe a statistically significant decline in test scores, with the typical student scoring 1.1 points lower in 2003 than in 1964 (significant at the 5 percent level). When we analyze boys and girls separately, the drop is of similar magnitude for each group and is statistically significant at the 10 percent level. Rescaling the scores to a common scale, the mean scores (for boys and girls combined) were 51.4 in 1964, 49.3 in 1978, 51.6 in 1994, and 50.2 in both 1999 and 2003. In 1999, the most recent year that the IEA surveys covered multiple grades,

5. The Rasch item response model is used to create a standardized score for the 1964, 1978, and 1994 tests. However, not all the common items in the 1964–94 tests also appear in the 1999 and 2003 tests. We therefore take advantage of the fact that for the 1994–2003 tests, the IEA has already created a variable called National Math Rasch Score. We standardize this variable so that in 1994 it has the same mean and standard deviation as the Rasch score derived from common items in the 1964–94 tests.

6. On the face of it, this approach differs from that used to analyze the LSAY data. If we estimate the equivalent of equation 1 with LSAY data using all observations with student age, grade, school sector, and state controls, the declines between the first and last cohorts for all students are 1.7 and 1.2 points for literacy and numeracy, respectively—smaller but close to the estimates in figures 1 and 2. These declines were statistically significant at the 1 percent level. The alternative methodologies do not produce results that differ qualitatively. Where we also address the potential endogeneity of age and grade in the equation using LSAY data (which is another way of characterizing Rothman's argument for using the whole year 9 cohort), the results do not change in any qualitative way. We use as instruments for age and grade the student's birth month and its interaction with school commencement rules. These rules, which vary between jurisdictions and changed over time, largely determine students' ages and grades when surveyed in the LSAY cohorts. The declines in performance between the first and last surveys were 1.4 and 0.8 points for literacy and numeracy, respectively, with the former significant at the 1 percent level and the latter at the 5 percent level.

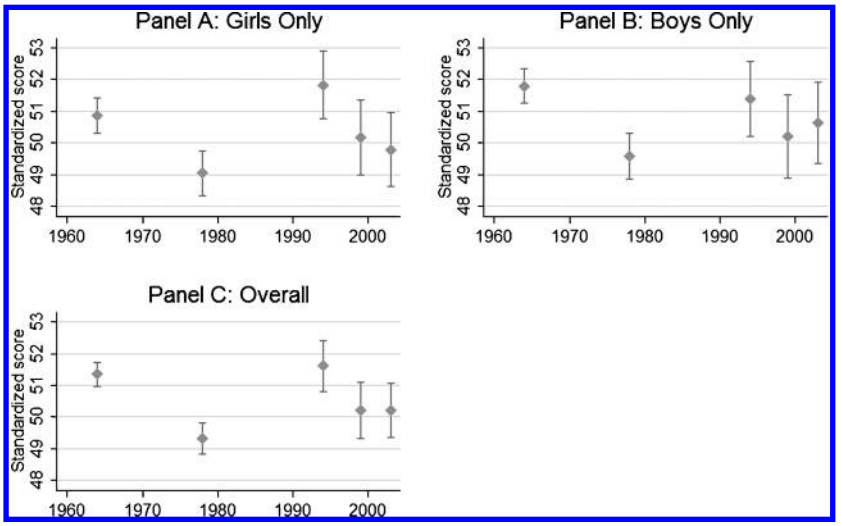

Figure 4. Numeracy (IEA). Notes: Effects are adjusted for student age and grade. Capped spikes depict 95% confidence interval for mean.

students' mathematics scores rose by an average of four points per grade. These findings therefore imply that the numeracy of the typical young teenage student in 2003 was approximately a quarter of a grade level behind his or her counterpart in 1964.

Although we use a somewhat different methodology, our results from 1964 to 1978 are close to those of Afrassa and Keeves (1999), who find a decline of approximately half a year of mathematics learning between 1964 and 1978. However, while Afrassa and Keeves find a decline of approximately one year of mathematics learning between 1964 and 1994, we observe no change in scores over this period. This is most likely because our results adjust for both age and grade effects, while Afrassa and Keeves adjust only for grade effects.

To see whether the fall in mean scores was driven by changes in particular points of the distribution (either across individuals or across schools), figure 5 presents kernel density plots of the distributions in 1964 and 2003. As with the LSAY density functions, the horizontal axis shows the standardized score, while the vertical axis shows the share of observations that are at or close to each score. Panel A shows the distribution of individual numeracy scores, indicating that the distribution has simply moved toward the left, with little change in dispersion. Panel B shows the distribution of average numeracy scores at the school level, with the distribution shifting to the left over time. In addition, there appears to be greater school-level dispersion. However, the apparent change in dispersion is probably due mostly to changes in the sampling frame. While the 2003 IEA test sampled schools across the nation,

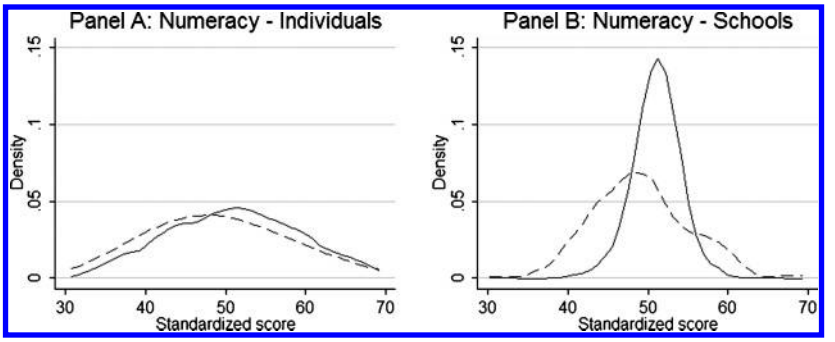

Figure 5. Distributions (IEA). Notes: Effects are adjusted for student age and grade. Continuous line represents 1964; dashed line represents 2000.

the 1964 test included only government schools in five states. The omission of nongovernment schools from the earlier test is the most likely reason why the school-level distribution appears narrower in 1964 than in 2003.

Recent Evidence—Has There Been a Turnaround?

Although our main results are based on the LSAY and IEA surveys, it is worth briefly noting the available evidence on trends in literacy and numeracy in Australia during the 1990s and 2000s. Reports by the state auditor-general's offices in New South Wales and Victoria (Australia's two largest states) found little change in literacy and numeracy scores over the periods 1996–2007 and 1998–2007, respectively (NSW Audit Office 2008; Victorian Auditor-General's Office 2009). The Programme for International Student Assessment (PISA 2007) provides comparable data on reading over the period 2000–2006 and on mathematics over the period 2003–6. According to PISA, the reading performance of Australian fifteen-year-olds fell from 2000 to 2006 (a statistically significant drop), and the mathematics scores of Australian fifteen-year-olds fell from 2003 to 2006 (statistically insignificant). The other comparable survey is the Adult Literacy and Life Skills Survey (ABS 2007), which found that the document literacy and prose literacy of 15–19-year-olds fell slightly from 1996 to 2006 (though not by a statistically significant margin). Over the same period, document and prose literacy in the general population rose slightly.

Other Evidence—Flynn Findings

It is also worth mentioning other sources of data on the cognitive abilities of Australian children that may (at least initially) appear to contradict the results presented here. The first is the well-known Flynn effect, under which IQ scores in many developed nations have shown an increase during the

Table 1. Long-Run Test Score Changes in Australia

Age	Test	Period	Total Change (LSAY scale)	Total Change (IQ scale)	Years	Change per year (IQ scale)
Panel A: Flynn findings (1936–81)						
10–14	Jenkins	1949–81	10.45	15.67	32	0.490
10–16	Ravens	1950–76	5.84	8.76	26	0.337
10–14	Otis	1936–49	3.67	5.50	13	0.423
Panel B: Trends in school tests (1964–2003)						
14	LSAY literacy	1975–98	−2.07	−3.10	23	−0.135
14	LSAY numeracy	1975–98	−1.67	−2.50	23	−0.109
13	IEA numeracy	1964–2003	−1.14	−1.71	39	−0.044

Notes: The LSAY scale has a standard deviation of 10. The IQ scale has a standard deviation of 15. All changes are statistically significant at the 5 percent level or better.
Source: Panel A is from Flynn (1987). Panel B is from authors' calculations.

twentieth century (Flynn 1987, 2006). If the test scores of young teenagers are stagnant, how can it be the case that IQ scores are rising in the general population?

There are three answers to this. First, in the case of Australia, Flynn's data are drawn from an earlier era. Table 1 juxtaposes Flynn's results with those shown above. We follow Flynn in converting the LSAY and International Mathematics tests into the IQ scale (with a mean of 100 and a standard deviation of 15). While Flynn (1987) demonstrated that Australian test scores grew by 0.4–0.5 points per year in the interwar period and the immediate postwar decades, the data shown above indicate that the last three decades of the twentieth century saw a slight decline in test scores.[7]

The second point to be made about Flynn's findings is that—across a wide range of countries—they appear to be more robust when tests are administered to adults rather than to younger children. Flynn (1987) presents strong data on test score gains for seven countries. In five of these countries, the data are for adult subjects. As he acknowledges, factors that might be driving higher test scores among the adult population are higher school completion rates and rising rates of postsecondary education. Even if test scores of schoolchildren were flat, more years of education might be expected to boost adult test scores.

7. See also de Lemos (1997), who discusses evidence on changes in Australian students' test scores from the 1950s to the 1990s. One contrary study is Nettlebeck and Wilson (2004), who found that the Peabody Picture Vocabulary Test scores of a small sample of Australian schoolchildren rose between 1981 and 2001. However, both their samples are drawn from a single primary school, making it quite possible that the observed changes were due to local demographic shifts and were not nationally representative.

The third factor to bear in mind when reconciling Flynn's findings with those presented here is that the Flynn effect is largest for standard IQ tests (such as the Ravens, Stanford-Binet, and Wechsler tests) and is virtually absent for curriculum-based tests. Discussing evidence for the United States and other developed countries, Flynn (1999, p. 8) noted that long-run test score gains are "small to nil on achievement tests. That is to say, they fall away the closer we come to the content of school-taught subjects." Indeed, Flynn (2009) presented evidence that in Britain, test scores for children aged 14–15 declined slightly over the period 1980–2008.

Together, these three factors—different time periods, different ages, and different types of tests—suggest that the absence of a discernable increase in Australian test scores of young teenagers on a curriculum-based test is readily reconcilable with the available evidence on the Flynn effect.

Other Evidence—International Comparisons

The other piece of evidence that may (to some eyes) appear to contradict the long-run trends is the fact that Australia ranks highly when compared with other developed countries. Pooling data for three international tests—the 1995, 1999, and 2003 TIMSS tests, the 2000 and 2003 PISA tests, and the 1994–98 International Adult Literacy Survey (IALS)—Brown et al. (2005) compare rankings for a group of eighteen Organisation for Economic Co-operation and Development (OECD) countries.[8] This comparison has the advantage that all the countries are of a similarly high income level (on a purchasing power parity basis, Australia's gross domestic product per capita is ranked ninth among these eighteen countries).

Pooling these data sources, Brown et al. find that the ranking of the median Australian respondent is on average 6.6 out of 18, indicating that the literacy, numeracy, and scientific knowledge of young Australians is somewhat above the OECD average. For PISA reading and for TIMSS and PISA mathematics, the typical Australian student is two-thirds to three-quarters of a standard deviation above the OECD mean. For PISA science and TIMSS science, the typical Australian student is nearly a full standard deviation above the OECD mean, and for the IALS literacy tests, the typical Australian student is at or slightly below the OECD mean.

Just as a rich country can experience a period of no economic growth yet still remain better off than many other countries, Australia's above-average ranking on international tests can easily be reconciled with the failure of test scores to rise over recent decades. Countries that rank reasonably well

8. Although IALS covers the entire adult population, Brown et al. (2005) analyze only respondents aged 16–24.

in international comparisons are not necessarily also improving over time. Indeed, the findings of Gundlach, Woessmann, and Gmelin (2001) indicate that test scores in OECD countries were essentially flat over the period 1970–94. It is therefore unsurprising that the failure of Australian test scores to rise has not dragged the nation to the bottom of the international league tables.

3. ALTERNATIVE EXPLANATIONS

In this section, we canvass four possible explanations for the fact that test scores have not risen over time. Perhaps student demographics have changed in such a way that we would have expected the more recent cohort of students to under-perform earlier cohorts. Maybe social factors, such as bullying and television viewing, have made children harder to teach. Perhaps lower-ability students dropped out of school in the 1960s and 1970s. Or maybe those literacy and numeracy questions that are comparable over time are the most irrelevant to the modern labor market.

Demographic Shifts

One plausible explanation for the decline in test scores over time is that the demographic characteristics of students have changed in such a way as to affect test scores. For example, it tends to be the case that students from a language background other than English do worse on literacy and numeracy tests (see, e.g., Rothman 2002). Since the share of students has risen over time, this would be expected to reduce the average test scores. Conversely, it tends to be the case that students with university-educated parents do better on literacy and numeracy tests (see, e.g., Cardak and Ryan 2006). Because the share of students with university-educated parents has risen over time, this would be expected to raise average test scores. Thus the overall effect of changing demographics is ambiguous.

To separate the effect of changing demographics, we employ a technique known as Oaxaca decomposition (Oaxaca 1973). Although traditionally used to decompose gender and racial pay differences, Oaxaca decompositions have also been employed to look at test scores (see, e.g., Cook and Evans 2000). In the present case, such a decomposition allows us to separate the change in test scores into three components: changing attributes of the student body, changing returns to these attributes, and shifts that are not explained by demographics. However, it is important to note that such an exercise is limited by the fact that our data sets do not contain a comprehensive set of demographic characteristics. We return to this issue below.

In the case of comparing LSAY test scores from 1975 and 1998, the Oaxaca decomposition involves estimating the following regressions, where t denotes the test score, Z_j is a vector of student demographics, α is a constant, β_j are

estimated parameters, and ε is a normally distributed mean-zero error term. Subscript i indexes individuals, j indexes demographics, and 75 and 98 denote the 1975 and 1998 tests, respectively:

$$t_{i75} = \alpha_{75} + \beta_{j75} Z_{ij75} + \varepsilon_{i75}$$
$$t_{i98} = \alpha_{98} + \beta_{j98} Z_{ij98} + \varepsilon_{i98}.$$

Denoting mean scores in 1975 and 1998 as T_{75} and T_{98}, respectively, the change in mean test scores can be written as:

$$\Delta T = T_{98} - T_{75} = (\alpha_{98} + \beta_{j98} Z_{j98}) - (\alpha_{75} + \beta_{j75} Z_{j75}).$$

Equivalently,

$$\Delta T = \beta_{j98}(Z_{j98} - Z_{j75}) + Z_{j75}(\beta_{j98} - \beta_{j75}) + (\alpha_{98} - \alpha_{75}). \quad (2)$$

The first term on the right side of equation 2 is the change in test scores that can be attributed to shifts in student demographics, the second term is the component due to changes in the coefficients on those demographics, and the third term is the component that cannot be explained by demographics or coefficients.

To see this more precisely, note that if student demographics are exactly the same in 1975 and 1998, the first term should be zero. If the returns to student demographics are exactly the same in 1975 and 1998, the second term should be zero. And if the change from 1975 to 1998 is entirely explained by changes in demographics and the returns to those demographics, the third term should be zero.

In addition, it is important to note that the effects need not go in the same direction. That is, it is perfectly possible for the returns to student demographics to contribute to a rise in test scores and for an unexplained component to contribute to a fall in test scores.

Table 2 presents results. Among the sample of students for whom we have non-missing demographic characteristics in the LSAY, the decline in test scores from 1975 to 1998 is only around half as large as across the full population (−1 point for literacy and −0.6 points for numeracy).[9] However, when

9. The main reason for missing demographics in the LSAY is that a number of the relevant questions were not asked in the first wave. For example, we drop almost 20 percent of observations in the LSAY 1975 cohort (principally because questions on parental occupations were not asked until 1978) and 36 percent of observations in the LSAY 1998 cohort (principally because the question about siblings was not asked until 2000). To see how this might have affected our analysis, we focus on the dropped cases (those with some missing demographics) and check how the key demographic patterns in these cases shifted over time. We observe a similar pattern to that seen in table 2: among the dropped cases, there are increases in parental education levels, in the proportion of students

Table 2. Decomposition of Changes in Test Scores

	(1) LSAY Literacy (1975–98)	(2) LSAY Numeracy (1975–98)	(3) IEA Numeracy (1964–2003)
Change in average test score (all students in sample)	−2.07	−1.67	−1.14
Decomposition			
Change in average test score (students with non-missing demographics)	−1.00	−0.58	−0.90
Earlier year as base			
Change attributable to levels of student demographics	+1.44	+1.92	+3.64
Change attributable to returns to student demographics	+0.98	+1.36	+3.02
Change not attributable to observed demographics	−3.42	−3.86	−7.56
Later year as base			
Change attributable to levels of student demographics	+1.21	+1.19	+3.81
Change attributable to returns to student demographics	+1.21	+2.09	+2.85
Change not attributable to observed demographics	−3.42	−3.86	−7.56

Notes: LSAY: Sample is 14-year-olds in grade 9, and demographics are indicators for student gender, whether school is in a metropolitan area, whether it is a Catholic school, whether it is an independent school, whether student was born in a non-English-speaking country, whether student's mother was born in a non-English-speaking country, share of students in the school with a mother born in a non-English-speaking country, number of siblings the student has, whether student's father has a degree, whether student's mother has a degree, parents' occupational status (on the ANU 3 scale), and indicators for each state and territory. IEA: Test scores are the residual from a regression of the Rasch score on indicators for age and grade. Demographics are student gender, an indicator for whether student/school is in a metropolitan area, and indicators for whether student's mother and father have lower secondary education, upper secondary education, a trade qualification, or a university degree.

we decompose the change, we find that shifts in demographic characteristics should have acted to increase test scores over time. Taking account of shifts in the levels of and returns to student demographics, the decline in test scores is considerably larger: around 3.4 points for literacy and 3.9 points for numeracy.

For the IEA test, our approach is very similar, except that because the age and grade coverage of the test changes over time, we take as the dependent variable the residual from a regression of the student's test score on indicator variables for age and grade. This can be thought of as parsing out differences in test coverage. The Oaxaca decomposition then compares the 1964 and 2003

from non-English-speaking backgrounds, and in the proportion attending private schools. This suggests that dropping cases with missing demographics is unlikely to have significantly biased the LSAY analysis in table 2. (Unfortunately we cannot carry out a similar robustness check for the IEA because that analysis focuses principally on parental education, which is missing for about 44 percent of cases in the 1964 survey and about 38 percent of cases in the 2003 survey.)

test cohorts. Among the sample of students for whom we have non-missing demographic characteristics in the IEA, the decline in test scores from 1964 to 2003 is similar to the decline in the full sample. When we decompose the shift, we find that shifts in demographic characteristics should have acted to increase test scores over time. Taking account of shifts in the levels of and returns to student demographics, the decline in test scores is 7.6 points. However, this is partly a function of the very limited demographic characteristics available in the IEA tests. If we had the same demographic variables in the IEA as we have for the LSAY, we expect that the adjusted decline in test scores would be smaller than 7.6 points.

In interpreting the results of this decomposition, it is important to note that in both data sets the demographic variables available to us are less extensive than one might like. In particular, we lack precise information on whether the child is in a single-parent family, and our data on neighborhood characteristics are not as precise as would be ideal.[10]

Social Trends

Another plausible explanation for the failure of test scores to rise over time is that major social trends adversely affected the school performance of Australian children. For example, if school violence and bullying had prevented students from learning, or television viewing had undercut children's ability to study at home, it is possible that such factors might have masked an underlying rise in school productivity. In theory, the effect of such social trends would be analogous to the impact of changing demographics, except that the offsetting factors considered here are behavioral and technological rather than demographic.

One possible social factor that might have lowered test scores is if there had been an exogenous increase in violence in schools. Such an increase would have made the job of schools more difficult, potentially counteracting other factors that would have caused scores to rise. To test this, we would ideally have liked comparable data on school-related violence that were unaffected by changes in punishment regimes. While we were unable to obtain such data, our analysis of the literature on youth crime rates in Australia suggests that violent and property crime rates have stayed constant through the 1980s, 1990s, and 2000s (see, e.g., Wundersitz 1993, 2005; Carcach 1997; AIHW

10. When analyzing the impact of social trends (e.g., television viewing), we use only data from the 2003 IEA data set to place an upper bound on the impact of these factors. In theory, such an approach could also have been used to put a bound on the impact of parental employment or family structure, but these demographic variables are unfortunately not available in the 2003 IEA.

2007).[11] It therefore seems unlikely that school-related violence has increased markedly over this period.

Another possibility is that increased television viewing by children made the job of schools more difficult. This could occur either because television viewing crowded out time that would have otherwise been spent doing homework or because children who watch more television have a shorter attention span.[12] The only survey of children's television viewing habits that we have been able to locate for the 1960s is a 1960 survey of Perth households, which found that children watched an average of one hour per day.[13] By contrast, IEA data for 2003 suggest that children watched around two hours of television per day. This extra hour per day of television viewing could plausibly have had an adverse impact on test scores.

A third social factor that might have affected test scores is attitudes to learning. Anecdotally, teachers sometimes report that the current cohort of children is more difficult to teach than cohorts from the 1960s and 1970s. It is sometimes said that children are now less inclined to appreciate school as a learning environment than they were in the past. Although we do not have comparable data on these social measures in the early test score instruments, proxies for some of them appear in the 2003 IEA survey. Therefore one way to put bounds on the impact of these social trends is to look at our most recent IEA survey—assuming that the 2003 coefficients can be applied to other periods—to see how much a worsening in these variables could have affected test score trends.

For simplicity, we restrict our analysis to three variables: whether the student agrees with the statement "I like being in school," whether the student says that in the past month "I was hit or hurt by other student(s) (e.g., shoving, hitting, kicking)," and the number of hours of television and videos that the student watches on a normal school day.

Table 3 shows the results of this exercise. Without parental demographics, students who like school score 2 points higher (statistically significant), students who are bullied score 0.3 points lower (not statistically significant), and an additional hour of television watching is associated with test scores that are 0.8 points lower (statistically significant). These are relatively small effects,

11. Wallace (1986) finds a rise in homicides by offenders aged 15–19 between 1958–67 and 1968–81. However, this is partly explained by the higher share of the population in this age band in the latter period.
12. Using a convincing empirical strategy arising from the quasi-random rollout of television across the United States, Gentzkow and Shapiro (2008) find no evidence that early exposure to television affected children's subsequent test scores.
13. Carter (2005) found that in a sample of 203 Perth children, 86 had a television in the home. Among those with a television in the home, the average daily viewing was 2 hours and 18 minutes, making the mean television viewing time across all children about 1 hour per day.

Table 3. Effect of Social Factors on Numeracy Performance (2003)

	(1)	(2)
Likes school (indicator)	1.924	1.041
Mean = 74%	[0.458]	[0.444]
Bullied (indicator)	−0.325	−0.267
Mean = 27%	[0.462]	[0.441]
TV watching (hours per day)	−0.840	−0.549
Mean = 1.96 hours	[0.168]	[0.165]
Student age and gender controls	Yes	Yes
Parental SES controls	No	Yes
Observations	3683	3683
R^2	0.05	0.15
Predicted impact on scores if:		
All children like school	+0.500	+0.271
No children are bullied	+0.088	+0.072
No children watch TV	+1.646	+1.076
All of the above	+2.234	+1.419

Note: Robust standard errors are in brackets. Both regressions include indicators for student gender and student age (in months). Parental socioeconomic status (SES) controls are whether school is in a metropolitan area and includes indicators for number of books in the home, number of people in the home, how often English is spoken at home, mother and father's educational attainment, and whether mother and father were born overseas. Sample is respondents from the Australian wave of the 2003 IEA, with non-missing demographics.

given that the standard deviation is 10 points. Including parental demographics in the regression reduces the magnitude of all three coefficients.

These results allow us to predict the impact on scores if (1) every child likes school, (2) no children are bullied, (3) no children watch television at home, and (4) all of the above. It is difficult to imagine that any of these are in fact an accurate characterization of Australian schoolchildren in 1964 (as noted above, one study found that children watched an average of one hour of television per day in 1960). Yet their combined effect is still only in the order of 1–2 points (0.1–0.2 standard deviations), which is barely more than the observed drop in IEA test scores from 1964 to 2003. At most, these three social factors could account for the decline in IEA test scores, but not for their failure to rise.

Just as the demographic analysis in the previous subsection has its limits, it is also worth acknowledging that this exercise cannot fully reject the possibility that social trends affected Australian test scores. For example, because the 2003 IEA data set does not contain questions about a child's diet, we cannot test the impact of changing nutrition on test scores. It is also possible that social factors affect test scores in a way other than we have modeled—for example, in

a highly nonlinear fashion, with strong interaction effects, or with important differential impacts across regions or schooling sectors.

Early School Leaving

One plausible explanation for the failure of test scores to rise over time is the possibility that in our earliest cohorts, a significant number of students dropped out of school before the test was conducted. Assuming that school leavers would have scored below average on the test, reduced rates of school leaving would lead to a decline in test scores. However, note that if all students obeyed the compulsory school-leaving laws, this should not have occurred. In the 1964 IEA, most students were thirteen years old, an age at which Australian children were required to attend school in all states and territories. In the 1975 LSAY, all students were fourteen years old, again an age at which Australian children were required to attend school in all parts of the country.

To test whether students actually complied with the school-leaving laws, we checked official statistics on the enrollment of the cohort that was to be thirteen-year-olds in 1964 (using the *Australian Yearbook* for various years). If early dropout was a problem, one would expect to see that the school enrollment of this cohort was larger in prior years. However, we observe little evidence of this. By comparison with the cohort of thirteen-year-olds in 1964, the cohort of twelve-year-olds in 1963 was 0.14 percent larger, the cohort of eleven-year-olds in 1962 was 0.02 percent larger, and the cohort of ten-year-olds in 1961 was actually 0.32 percent smaller. This kind of trivial variation suggests that in the early 1960s, very few children dropped out of school between their tenth and thirteenth birthdays.[14]

We conducted a similar exercise on the first LSAY cohort, comparing prior years' enrollment with the age cohort that was to turn fourteen in 1975 (unfortunately, we were not able to obtain enrolment statistics for 1974).[15]

14. Three other factors affecting this calculation are immigration, emigration, and death. The average net immigration rate (inflows minus outflows) was 0.75 percent in 1961–64 and 0.49 percent in 1971–75. In both periods, young teenagers were probably underrepresented in population movements. The annual probability of death for a person aged 10–15 in the 1960s and 1970s was around 0.03–0.04 percent. Both sets of figures indicate that these factors are unlikely to make a significant difference to our results.

15. To see whether the absence of 1974 enrollment data affected the results (and to avoid potential problems created by combining enrollment data from different data sources), we replicated the analysis using data from New South Wales, Australia's largest state. All data were obtained from the annual state yearbooks. By comparison with fourteen-year-olds in 1975, we found that the cohort of thirteen-year-olds in 1974 was 1.05 percent smaller, the cohort of 12-year-olds in 1973 was 2.29 percent smaller, the cohort of eleven-year-olds in 1972 was 2.48 percent smaller, and the cohort of ten-year-olds in 1971 was 2.50 percent smaller. This implies that in the early 1970s, about 2 percent of children in NSW dropped out of school between their tenth and fourteenth birthdays. Supposing that NSW dropouts would have scored 1 standard deviation below average, this would have biased the true result upward from 49.8 to 50.0.

By comparison with fourteen-year-olds in 1975, we found that the cohort of twelve-year-olds in 1973 was 1.26 percent smaller, the cohort of eleven-year-olds in 1972 was 1.27 percent smaller, and the cohort of ten-year-olds in 1971 was 0.68 percent smaller. This implies that in the early 1970s, about 1 percent of children dropped out of school between their tenth and fourteenth birthdays. To see the largest effect this attrition could have had on the mean test scores, suppose that those who dropped out of school would have scored a full standard deviation lower than those who remained. If the mean score of those who stayed had been 50, this implies that the true mean score would have been 49.9. Such variation would have had virtually no effect on the observed results.

Have 1960s and 1970s Questions become Irrelevant?

Another alternative explanation for the observed changes is that the questions asked in earlier years are simply irrelevant today. For example, if it were the case that literacy questions in the mid-1970s required a knowledge of words that had fallen out of common usage by the 1990s, or if mathematics questions in the mid-1960s required a level of mental arithmetic that many would judge to be unnecessary in the early 2000s, one might wonder whether it was reasonably possible to compare performances on the two tests.

Although there is not a comprehensive way to address this critique, one straightforward approach is to look at the tests themselves. The appendix, which can be accessed at the *Education Finance and Policy* Web site (http://www.mitpressjournals.org/loi/edfp), presents several examples of questions that were common across IEA tests. In our view, these common questions are reasonably representative of modern-day mathematics tests, although they are all computational rather than conceptual. It is therefore possible that the IEA comparison placed more emphasis on computational skills.

Marks and Ainley (1997) analyzed responses to the tests conducted in 1975 and 1995 as part of the LSAY series and found that there were counteracting tendencies in both the literacy and numeracy tests. For the literacy tests, the report found that students in the later cohort were more likely to answer correctly questions relating to newspaper articles but were less likely to answer questions dealing with more difficult textual passages correctly. For the numeracy tests, Marks and Ainley found that students' performance was poorer on computational items in the later cohort but had improved on conceptual items.

4. EXPENDITURE TRENDS

In Australia, we have been unable to find a consistent series of school expenditures covering the period since 1964.[16] We therefore combine a number of different data sources to produce a consistent series of school expenditures, including both government and private spending. Government expenditure is obtained from various official tabulations and includes expenditures from all levels of government provided to both public and private schools. Private expenditure is calculated from household expenditure surveys, from which we estimate private spending as a share of government spending. Taking account of changes in the share of children attending public and private schools, we then estimate private school spending in all years. The online appendix provides details of the derivation of our series.

One possible limitation to this estimate is that we are unable to disaggregate spending into primary and secondary school expenditure. Since per pupil expenditure is typically higher for secondary school pupils than for primary school pupils, and since rising school completion rates mean that the share of pupils in secondary schools is higher in the early 2000s than in the mid-1960s, this will bias upward the trend in school spending. Empirically, however, the extent of this bias is very small. In 2002–3, per student spending in government schools was 28 percent higher at the secondary school level than at the primary school level (MCEETYA 2003, appendix 1: statistical annex, table 20). Over the period spanned by our tests—1975–98 and 1964–2003—the share of pupils in secondary school grew by 4.6 and 4.4 percent, respectively.[17] This means that the upward bias to school spending caused by an increase in the share of students at the secondary school level was only 1.3 percent.

To adjust for price changes, we index the expenditure series in three ways. First, we adjust using the All Groups Consumer Price Index (CPI), which assumes that schools' input prices rose at the same rate as other goods and services in the economy. Second, we construct a Schools Price Index (an index of the prices of inputs used by schools) to account for the possibility that the price of schools' inputs has been growing at a different rate from other prices. And third, we construct a price index based on the earnings of professional women, on the assumption that the largest expenditure item for schools is

16. In their study of schooling productivity in OECD countries over the period 1970–94, Gundlach, Woessmann, and Gmelin (1999, appendix) used education expenditure data reported to the United Nations Educational, Scientific, and Cultural Organization (UNESCO).

17. An eagle-eyed reader may wonder at the fact that the secondary share grew more over the period 1975–98 than over the period 1964–2003. It is important to remember that the figure is affected not only by school completion rates (which have risen steadily) but also by cohort-specific factors, such as the baby boom.

Table 4. Increase in Schools Expenditure (percent) (preferred estimates in bold text)

Year	Nominal Spending per Child	Indexed by All Groups Price Index	Indexed by Schools Price Index	Indexed by Earnings of Professional Women
Panel A: Private and government expenditure				
1974–75 to **1997–98**	429%	18%	**10%**	−2%
1963–64 to **2002–3**	4,147%	333%	**258%**	76%
Panel B: Government expenditure only				
1974–75 to **1997–98**	410%	14%	6%	−5%
1963–64 to **2002–3**	3,962%	314%	243%	69%

teacher salaries, and around 60–70 percent of teachers were women in this period.[18]

Panel A of table 4 sets out the proportionate change in total school spending over time. Using the All Groups CPI, spending increased by 18 percent over the period of the LSAY and by 333 percent over the period of the IEA tests. Using the Schools Price Index, spending increased by 10 percent over the period of the LSAY tests and by 258 percent over the period of the IEA tests. Indexing schools' expenditures by the earnings of professional women, spending declined by 2 percent over the period of the LSAY and rose by 76 percent over the period of the IEA tests. The substantially lower expenditure growth rates over the period spanning the LSAY data reflect the strong spike in per capita expenditure just prior to 1975, arising from school expenditure decisions made by the Whitlam government on the advice of the Interim Committee (1973) report (the Karmel report). This spike is captured in the IEA growth rates but not in the growth over the period spanned by the LSAY.

With all three price indices, real spending rose over the period of the IEA tests. Using two of the three price indices, real spending rose over the period of the LSAY tests. The exception is when spending is indexed by the earnings of professional women. This is consistent with Leigh and Ryan (2008), who find that—relative to professionals—new teachers' earnings declined from the mid-1970s to the mid-2000s.[19] However, this index almost certainly overstates the expenditure that would have been necessary to keep pace with female

18. To be precise, the share of teachers who were women was 67.4 percent in 2003, 65.9 percent in 1998, and 59.5 percent in 1975. This calculation includes both primary and secondary schools and the government and nongovernment sectors. We were unable to find the share in 1964, but the share of female teachers was 59.7 percent in 1970 and 55.0 percent in 1950. Linearly interpolating, this suggests that the share was 58.3 percent in 1964.

19. The drop in teacher pay relative to other professionals is not confined to new teachers. A similar decline is also evident if one compares earnings for all teachers with those of all professionals (adjusted for the age composition of both groups). See Leigh and Ryan (2006, appendix III).

professionals' wages (and therefore understates the rise in real spending). In 2002–3, teacher salaries comprised 63 percent of government school expenditures.[20] Even if one thought that teachers' salaries should have risen at the same rate as professional women's salaries, it would be more appropriate to take a weighted average of the All Groups price index and the Earnings of Professional Women price index. This approach implies a slightly lower rise than the Schools Price Index. (Using weights of 37 percent for the All Groups price index and 63 percent for the Earnings of Professional Women price index, this approach implies a 5 percent rise in real education expenditures between 1975 and 1998 and a 171 percent rise over the period from 1964 to 2003.)

Since we have more precise data on government expenditure than on private expenditure, Panel B of table 4 replicates the results using only government expenditure. Over the periods covered by the LSAY and IEA tests, government expenditure on schools has risen substantially, though not as rapidly as private expenditure. Again, if expenditure is indexed by the earnings of professional women, we observe a decline over the period 1975–98. However, a weighted average of the All Groups price index and the Earnings of Professional Women price index still implies an increase in government school expenditure over this period.

Our preferred estimate uses both private and government expenditure, on the basis that our test scores are for children in both government and nongovernment schools. Our preferred price index is the Schools Price Index. We prefer this to the All Groups Price Index because it allows for the cost of educational inputs to rise more rapidly than the prices of other items in the economy. We also prefer it to the Earnings of Professional Women price index because the Schools Price Index does not make the unrealistic assumption that salaries constitute 100 percent of school expenditure. Using both private and government expenditure, indexed to the Schools Price Index, suggests that real per child school expenditure increased by 10 percent over the period of the LSAY tests and by 258 percent over the period of the IEA tests.

Where has the increased educational expenditure gone? One significant factor pushing up costs over this period has been smaller class sizes. In the online appendix, we show student-teacher ratios for 1964–2004. On average, class sizes fell by 20 percent in the period 1975–98 and by 43 percent in the period 1964–2003. These patterns can be directly observed in our data: in

20. We calculate this figure using data in MCEETYA (2003, Appendix 1: statistical annex, table 19). Since the user cost of capital is not included in our estimates, we omit it when calculating the salary share of Australian government school expenditure.

the 1964 numeracy test, the average class size was 36, whereas in 2003, the average class size was 26.[21]

Assuming that salaries constitute 63 percent of all school spending, and that cutting class size requires a proportionate increase in funding, these cuts in class size would have boosted school spending by 13 percent over the period 1975–98 and by 27 percent over the period 1964–2003. Lower student-teacher ratios could therefore account for almost all of the expenditure increase over the period 1975–98, though not over the period 1964–2003.

One factor that might have increased expenditure without improving educational outcomes was the declining number of teachers in Catholic schools (and some independent schools) who were from religious orders, such as brothers and nuns. These teachers received a stipend that was less than the wage received by teachers in government schools at the time. As their numbers fell from the 1960s onward, they were replaced by lay teachers who were paid a similar wage to government schoolteachers. (Or in some cases religious order teachers received a salary increase to put them on a similar footing to government schoolteachers.) Unfortunately, data on the gap between the stipend paid to religious teachers and the salary paid to lay teachers are limited.[22] However, we do have data on the share of religious teachers in Catholic schools in various years.[23] Accounting for the share of students taught in Catholic schools, and assuming that wages constitute 63 percent of total expenditure, we can see how this factor would have affected total costs on the assumptions that the stipend was 25 percent, 50 percent, or 75 percent of a lay teacher's salary.[24] From 1964 to 2003, the increase in spending caused by the shift from religious to lay teachers would have led to a 7 percent spending increase if the stipend for religious teachers was 25 percent of lay salaries, 5 percent if the stipend was 50 percent of lay salaries, and 2 percent if the stipend was 75

21. Both the 1964 and 2003 IEA numeracy surveys contain data on average class size in the student's grade and school. While the bands are not directly comparable, they give some indication of the class size reduction over this period. In 1964, 7 percent of Australian students were in a class of twenty-four or fewer students, and 70 percent of students were in a classroom with thirty-five or more students. In 2003, 31 percent of students were in a class of twenty-four or fewer students, and four percent were in a class of thirty-three or more students. There is no systematic relationship in either survey between test scores and class size, though it is quite possible that nonrandom sorting might offset causal impacts in either direction.
22. The only estimate we were able to obtain was from the South Australian Catholic Education Office, which informed us that their current religious stipend equates to 62 percent of the average salary paid to Australian schoolteachers. We are grateful to Geoff Hallion of Catholic Education South Australia for providing us with this information.
23. Flynn and Mok (2002) provide data on the share of Catholic teachers from religious orders in 1965 (72.3 percent), 1971 (47.8 percent), 1993 (4.0 percent), and 2000 (1.6 percent). We linearly interpolate and extrapolate to obtain estimates for the years 1964, 1975, 1998, and 2003.
24. The share of students in Catholic schools in our years of interest was 19.6 percent (1964), 17.0 percent (1975), 19.7 percent (1998), and 19.9 percent (2003). Our estimates account for changes in the "market share" of Catholic schools.

percent of lay salaries. Over the period 1975–98, the impact on total spending would have been 4 percent, 3 percent, and 1 percent, respectively. While this is clearly nontrivial, the movement away from religious teachers in Catholic schools can account for only a small share of the full increase in spending over either period.

One way of estimating the change in productivity of Australian schools is to divide the test score trends by the expenditure trends. In effect, this exercise estimates the necessary expenditure required for each point on the literacy and numeracy tests. We find that school productivity has declined by 12–13 percent between 1975 and 1998 and by 73 percent between 1964 and 2003.[25] This contrasts starkly with multifactor productivity across the economy, which rose by 34 percent in the period 1975–98 and by 64 percent in the period 1964–2003 (ABS 2006).

5. DISCUSSION AND CONCLUSION

Depending on which measure we employ, the test scores of young Australian teenagers appear to have fallen slightly from the mid-1960s to the early 2000s. Although our confidence intervals in some cases include zero, we can reject the hypothesis that there has been a statistically significant increase in test scores over the past four decades. While we do not have as full a set of demographics as we would like, adjusting test scores to account for changes in the available demographic characteristics does not explain the decline; rather, the decline seems even larger. The results are economically significant, implying a drop of between one-fifth and one-tenth of a standard deviation. This suggests, for example, that the numeracy of the typical young teenage student in 2003 was approximately a quarter of a grade level behind his or her counterpart in 1964. This drop contrasts with the United States, where standardized test scores merely flatlined from the 1970s to the 1990s (Hanushek 1997; Hoxby 2002).

Turning to expenditure data, our preferred estimate is that real per child school expenditure increased by 10 percent over the period of the LSAY tests (1975–98) and by 258 percent over the period of the IEA tests (1964–2003). These results are not particularly sensitive to the exclusion of private expenditure or the use of other reasonable price indices. Dividing the test score trends by the expenditure trends suggests that school productivity declined by 12–13 percent over the period covered by the LSAY tests and by 73 percent over the period covered by the IEA tests. This is consistent with results from other

25. An alternative (narrower) measure of the increase in school inputs would be the reduction in student-teacher ratios, which fell by 20 percent over the period 1975–98, and by 43 percent over the period 1964–2003 (online appendix). If this were regarded as the only input, it would imply a reduction of school productivity in the order of 19–20 percent between 1975 and 1998 and a reduction of 32 percent between 1964 and 2003.

developed nations, and it suggests that resources alone are not the answer to improving school performance. Instead, education policy makers should rigorously evaluate the impact of new reforms and focus on raising the quality of education expenditure.

Although we have done our best to adjust our results given the data available to us, we cannot fully rule out explanations that have nothing to do with school productivity. For example, it is plausible that changing family structure, social norms, and entertainment media may have affected test scores. Alternatively, it might be the case that our measure of school outputs is unduly narrow and fails to capture factors that do not appear on our tests—such as physical fitness, critical thinking skills, problem solving skills, communication skills, or knowledge of science. Since we do not have data on these other outputs, it is not possible to know whether they have risen or fallen over time.

However, it is possible to identify explanations that relate directly to school productivity. First, part of the spending increase over the period in question resulted from class size cuts. If smaller classes have little or no impact on test scores (Hanushek 1998; Hoxby 2000; Rockoff 2009; but compare with Krueger 1999, 2003), this policy change would have led to a reduction in school productivity. A second possibility is that falling teacher quality led to a disproportionate drop in student performance. If a 10 percent reduction in real teacher salaries reduces student performance by more than 10 percent, falling teacher salaries could lower school productivity. A third productivity-related explanation is that shifts in the way schools were managed led to a decline in school productivity. For example, Australian universities and teachers' colleges shifted toward a whole-language approach to teaching reading in the 1970s (van Kraayenoord and Paris 1994). To the extent that this was less effective than other methods of reading instruction, it might also have led to a decline in school productivity.[26] Although our approach does not allow us to distinguish between these explanations, we hope it will encourage further analysis.

This article was originally prepared as a report for the Australian Department of Education, Employment and Workplace Relations (DEEWR). The views expressed are those of the authors and do not necessarily represent the views of the DEEWR. The authors thank Sheldon Rothman of the Australian Council for Educational Research for providing one set of the school achievement scales used in this article. He bears no responsibility for our use of those scales. We have benefited from comments from

26. Similarly, it is possible that the mainstreaming of students with disabilities reduced mean test scores. This might occur for two reasons. The first is if the positive peer effects for disabled students were lower than the negative peer effects for nondisabled students. We are not aware of any research that bears directly on this issue. The second is if a significant proportion of very low-scoring students were added to the group taking the tests. None of the individual distributions in figures 3 and 5 show such an addition. Rather, the decline in performance occurred across the entire distributions.

editor David Figlio, an anonymous referee, and seminar participants at the Ministerial Council on Education, Employment, Training and Youth Affairs Performance Measurement and Reporting Taskforce, and the Queensland Studies Authority Senior Schooling Conference. In the revision of this report for publication, Susanne Schmidt and Elena Varganova provided outstanding research assistance.

REFERENCES

Afrassa, Tilahun, and John Keeves. 1999. Changes in students' mathematics achievement in Australian lower secondary schools over time. *International Education Journal* 1(1): 1–21.

Atkinson, Anthony B. 2005. *Atkinson review: Final report—measurement of government output and productivity for the national accounts*. London: Palgrave Macmillan.

Australian Bureau of Statistics (ABS). 2006. *Australian system of national accounts, 2005–06*. Cat. No. 5204.0. Canberra: ABS.

Australian Bureau of Statistics (ABS). 2007. *Adult literacy and life skills survey, summary results 2006*. Cat. No. 4228.0. Canberra: ABS.

Australian Institute of Health and Welfare (AIHW). 2007. *Juvenile justice in Australia 2004–05*. AIHW Cat. No. JUV #2. Canberra: AIHW.

Boyle, Richard. 2006. Measuring public sector productivity: Lessons from international experience. Committee for Public Management Research Discussion Paper No. 35, Institute of Public Administration.

Brown, Giorgina, John Micklewright, Sylke V. Schnepf, and Robert Waldmann. 2005. Cross-national surveys of learning achievement: How robust are the findings? Institute for the Study of Labor Discussion Paper No. 1652.

Carcach, Carlos. 1997. Youth as victims and offenders of homicide. Paper presented at the Juvenile Crime and Juvenile Justice: Toward 2000 and Beyond Conference, Adelaide, Australia, June.

Cardak, Buly A., and Chris Ryan. 2006. Why are high ability individuals from poor backgrounds under-represented at university? Discussion Paper No. A06.04, La Trobe University.

Carter, Owen. 2005. Changes in obesity, sedentary behaviours and Perth children's television viewing from 1960 to 2003. *Australian and New Zealand Journal of Public Health* 29(2): 187–88.

Cook, Michael D., and William N. Evans. 2000. Families or schools? Explaining the convergence in white and black academic performance. *Journal of Labor Economics* 18: 729–54.

de Lemos, Marion M. 1997. Reading standards up or down: What do the test norms say? Paper presented at the Australian Association for Research in Education Annual Conference, Brisbane, March.

Douglas, James. 2006. Measurement of public sector output and productivity. Policy Perspectives Paper No. 06/09. Wellington: New Zealand Treasury.

Flynn, James R. 1987. Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin* 101: 171–91.

Flynn, James R. 1999. Searching for justice: The discovery of IQ gains over time. *American Psychologist* 54(1): 5–20.

Flynn, James R. 2006. Efeito Flynn: Repensando a inteligência e seus efeitos [The Flynn effect: Rethinking intelligence and what affects it]. In *Introdução à psicologia das diferenças individuais* [*Introduction to the psychology of individual differences*], edited by Carmen Flores-Mendoza and Roberto Colom, pp. 387–411. Porto Alegre, Brazil: ArtMed.

Flynn, James R. 2009. Requiem for nutrition as the cause of IQ gains: Raven's gains in Britain 1938–2008. *Economics and Human Biology* 7(1): 18–27.

Flynn, Marcellin, and Magdalena Mok. 2002. *Catholic schools 2000: A longitudinal study of year 12 students in Catholic schools 1972–1982–1990–1998*. Sydney: Catholic Education Office.

Gentzkow, Matthew, and Jesse M. Shapiro. 2008. Preschool television viewing and adolescent test scores: Historical evidence from the Coleman study. *Quarterly Journal of Economics* 123(1): 279–323.

Gundlach, Erich, Ludger Woessmann, and Jens Gmelin. 1999. The decline of schooling productivity in OECD countries. Kiel Institute of World Economics Working Paper No. 926.

Gundlach, Erich, Ludger Woessmann, and Jens Gmelin. 2001. The decline of schooling productivity in OECD countries. *Economic Journal* 111: C135–C147.

Hanushek, Eric A. 1997. The productivity collapse in schools. In *Developments in school finance, 1996*, edited by William Fowler Jr., pp. 183–95. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Hanushek, Eric A. 1998. The evidence on class size. W. Allen Wallis Institute of Political Economy Occasional Paper No. 98-1, University of Rochester.

Hanushek, Eric A., and Dennis D. Kimko. 2000. Schooling, labor-force quality, and the growth of nations. *American Economic Review* 90(5): 1184–1208.

Hoxby, Caroline M. 2000. The effects of class size on student achievement: New evidence from population variation. *Quarterly Journal of Economics* 115(4): 1239–85.

Hoxby, Caroline M. 2002. School choice and school productivity (or could school choice be a tide that lifts all boats?). NBER Working Paper No. 8873.

Interim Committee for the Australian Schools Commission. 1973. *Schools in Australia: Report of the Interim Committee for the Australian Schools Commission*. Canberra: Australian Government Publishing Service.

Jamison, Eliot A., Dean T. Jamison, and Eric A. Hanushek. 2006. The effects of education quality on income growth and mortality decline. NBER Working Paper No. 12652.

Krueger, Alan B. 1999. Experimental evidence of education production functions. *Quarterly Journal of Economics* 114(2): 497–532.

Krueger, Alan B. 2003. Economic considerations and class size. *Economic Journal* 113(485): F34–F63.

Leigh, Andrew, and Chris Ryan. 2006. How and why has teacher quality changed in Australia? Centre for Economic Policy Research Discussion Paper No. 534, Australian National University.

Leigh, Andrew, and Chris Ryan. 2008. How and why has teacher quality changed in Australia? *Australian Economic Review* 41(2): 141–59.

Marks, Gary, and John Ainley. 1997. Reading comprehension and numeracy among junior secondary school students in Australia. LSAY Research Report No. 3.

Marks, Gary, and Nicole Fleming. 1998a. Factors influencing youth unemployment in Australia 1980–1994. LSAY Research Report No. 7.

Marks, Gary, and Nicole Fleming. 1998b. Youth earnings in Australia 1980–1994: A comparison of three youth cohorts. LSAY Research Report No. 8.

Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA). 2003. *National report on schooling in Australia, 2003*. Carlton South, Australia: MCEETYA.

Nettlebeck, Theodore John, and Carlene Wilson. 2004. The Flynn effect: Smarter not faster. *Intelligence* 32: 85–93.

New South Wales (NSW) Audit Office. 2008. *Improving literacy and numeracy in NSW public schools: Department of Education and Training*. Sydney: NSW Audit Office.

Oaxaca, Ronald. 1973. Male-female wage differentials in urban labor markets. *International Economic Review* 14(3): 693–709.

Programme for International Student Assessment (PISA). 2007. *PISA 2006 science competencies for tomorrow's world, Volume 2: Data*. Paris: Organisation for Economic Co-operation and Development.

Rockoff, Jonah. 2009. Field experiments in class size from the early twentieth century. *Journal of Economic Perspectives* 23(4): 211–30.

Rosen, Ellen D. 1993. *Improving public sector productivity: Concepts and practice*. New York: Sage Publications.

Rothman, Sheldon. 2002. Achievement in literacy and numeracy by Australian 14-year-olds, 1975–1998. LSAY Research Report No. 29.

Ryan, Chris, and Louise Watson. 2004. Year 12 completion and retention in Australia in the 1990s. *Australian Journal of Labour Economics* 7(4): 481–500.

Steering Committee for the Review of Government Service Provision (SCRGSP). 2007. *Report on government services 2007*. Canberra: Productivity Commission.

United Kingdom. Department for Education and Skills. 2005. *Measuring government education output in the national accounts*. London: Department for Education and Skills, now known as the Department for Education.

van Kraayenoord, Christina E., and Scott G. Paris. 1994. Literacy instruction in Australian primary schools. *Reading Teacher* 48(3): 218–28.

Victorian Auditor-General's Office. 2009. *Literacy and numeracy achievement*. Parliamentary Paper No. 171, Session 2006–09. Victoria: Victorian Government Printer.

Wallace, Alison. 1986. *Homicide: The social reality*. Sydney: NSW Bureau of Crime Statistics and Research.

Weale, Martin. 2007. Following the Atkinson review: The quality of public sector output. *Economic and Labour Market Review* 1(7): 22–26.

Wundersitz, Joy. 1993. Some trends in officially recorded youth offending: A state-by-state comparison. In *National conference on juvenile justice proceedings No. 22*, edited by Lynn Atkinson and Sally-Anne Gerull, pp. 53–66. Canberra: Australian Institute of Criminology.

Wundersitz, Joy. 2005. *Juvenile justice in South Australia: A 2004 update*. Information Bulletin No. 47. Adelaide: South Australian Office of Crime Statistics and Research.