Editorial Manager(tm) for Educational Psychologist
Manuscript Draft

Manuscript Number: EP-D-09-00034R3

Title: Standardized testing and school accountability

Article Type: Scholarly Essay

Corresponding Author: Professor Dylan Wiliam,

Corresponding Author's Institution: Institute of Education, University of London

First Author: Dylan Wiliam

Order of Authors: Dylan Wiliam

Abstract: This article explores the use of standardized tests to hold schools accountable. The history of testing for accountability is reviewed, and it is shown that currently between-school differences account for less than ten percent of the variance in student scores, in part because the progress of individuals is small compared to the spread of achievement within an age cohort, and, possibly, due to lack of alignment between instruction and assessment. A review of the literature on the effects of the introduction of such tests in high-stakes accountability regimes suggests that the effects can be positive, and the size of the effects is substantial. Therefore, while the validity of such tests may be problematic in terms of the intended inferences, their introduction may nevertheless be justified by their impact. The paper concludes with a number of suggestions on improving tests for high-stakes accountability.

Standardized testing and school accountability

Dylan Wiliam

Institute of Education, University of London

Abstract

This article explores the use of standardized tests to hold schools accountable. The history of testing for accountability is reviewed, and it is shown that currently between-school differences account for less than ten percent of the variance in student scores, in part because the progress of individuals is small compared to the spread of achievement within an age cohort, and, possibly, due to lack of alignment between instruction and assessment. A review of the literature on the effects of the introduction of such tests in high-stakes accountability regimes suggests that the effects can be positive, and the size of the effects is substantial. Therefore, while the validity of such tests may be problematic in terms of the intended inferences, their introduction may nevertheless be justified by their impact. The paper concludes with a number of suggestions on improving tests for high-stakes accountability.

Introduction

Assessment is a key process in education. It is only through assessment that we can find out whether instruction has had its intended effect, because even the best-designed instruction cannot be guaranteed to be effective (see, for example, Denvir & Brown, 1986a, b).

One might expect, therefore, that assessment should be reasonably uncontroversial. All those with a stake in the outcomes of education—learners, teachers, parents, other taxpayers, employers, and the wider community—want to know what students have learned, and it seems plausible that this can easily be evaluated through the use of straightforward and familiar instruments, such as achievement tests.

However, as H. L. Mencken warned almost a century ago, "There is always an easy solution to every human problem: neat, plausible, and wrong." (Mencken, 1917). In this article, I explore the use of standardized tests in high-stakes accountability systems, and specifically, I argue that the systems currently in use have significant shortcomings that call into question some of the interpretations that are routinely based on the scores yielded by these tests.

Of course tests are designed for a variety of purposes, and their results are used in a variety of ways. One common classification of tests distinguishes between diagnostic, norm-referenced, and criterion-referenced tests. These terms are, of course, more

properly applied to the kinds of interpretations that are made on the basis of the test scores rather than to the test itself but the term "criterion-referenced test" can certainly be used as a shorthand term for a test designed to support criterion-referenced inferences. However, it is important to bear in mind that it is the inferences, rather than the test that are criterion-referenced.

This distinction is particularly important in view of the fact that test scores are often interpreted in ways that differ significantly from those intended by the designers of the test. Tests designed to support norm-referenced inferences about the level of achievement in, say, fourth grade mathematics, are used to make inferences about the extent to which students have mastered the fourth grade content standards for mathematics in their state. Conversely, tests designed to support criterion-referenced inferences about a student's mastery of fourth-grade content standards for mathematics are used to make inferences about the adequacy of education provided by the district.

My aim in this article is to take one specific application of testing—the use of standardized achievement tests for the purpose of holding teachers, schools and districts accountable—and to explore the extent to which the tests currently in use are able to support valid inferences about the quality of education provided.

I begin by discussing the idea of accountability, and then review briefly the history of testing for accountability within public schools. The key assumption of accountability testing is that differences in the achievement of students on standardized tests should

be primarily attributable to differences in the quality of education received by students, and so, in subsequent sections, I investigate the extent to which this is the case. I show that differences in the quality of schooling account for only a small proportion of the variation in student outcomes (in most countries, less that ten percent) primarily because the cumulative effect of differences in the average rate at which individual students learn results in a situation in which the progress of an individual student in a year is much smaller than the variability of achievement within an age cohort. Any inferences that differences in student scores on standardized tests therefore primarily reflect differences in quality of education received are therefore problematic. However, the evidence from comparisons between states within the US, and of comparisons of different national systems, suggests that high-stakes accountability systems can have a positive impact on student learning. The article therefore concludes with some reflections on how high-stakes accountability systems might be designed more effectively.

Accountability

The word "accountability" is used in a wide range of contexts, and has a number of different meanings. To be accountable can mean to be responsible, to be answerable, to be blameworthy, or even to be liable. However, the literal meaning of the term—that of being "held to account"—suggests there is an expectation that when a person, organization or entity is accountable, they can be expected or required to render an account of their actions (or inaction). The two immediate questions that follow are "to whom?" and "for what?" (Wescott, 1972; Bardach & Lesser, 1996).

If schools are to be accountable, to whom should they be accountable? The obvious answer is to those who pay for the provision of the service, and to those who consume it. It is common in much political debate within the United States to assume that this means taxpayers (who pay for the service) and parents (who are generally regarded, especially within discourses of school choice, as being the consumers), although of course these are often the same people. Further removed, but still regarded as having a legitimate stake in the outcomes of education, are employers, educational institutions subsequently attended by students, and, in recent years, the students themselves. However, when education fails, the social and financial costs are borne by the whole of society. Even retired people who earn too little to pay tax will bear the costs of failures in the education system, through increased crime (Levin *et al*., 2007; Carneiro, Crawford & Goodman, 2007) and lower levels of engagement in citizenship and other forms of "pro-social" behavior (Feinstein, Budge, Vorhaus & Duckworth, 2008). However, while teachers, schools and districts should therefore be accountable to all those who have a stake in society, it is still necessary for some agency to take responsibility for the design of the accountability system, for reviewing the information it produces, and for taking any necessary action.

A brief history of accountability testing

There is nothing new in the idea that results of simple testing procedures could be used to hold students and their teachers to account. Up to the end of the first third of the 19th century, public schools in England and Wales had been financed by voluntary

(and in general, religious) organizations. Between 1833 and 1853, the role of the state in funding education was expanded greatly, through the introduction of grants for the erection of new buildings, for the training of teachers, and for the encouragement of attendance, first in rural schools, and then in all schools. In 1858, a Royal Commission was set up, under the chairmanship of the Duke of Newcastle, "to inquire into the state of popular education in England and to consider what measures were required for the extension of sound and cheap instruction to all classes". The Commission's report, published in 1861, recommended, among other things, that the amount of public money paid to each elementary school should depend on three factors: the condition of the school buildings, student attendance, and the performance of the students attending the school in an oral examination, undertaken by one of the national school inspectors, of every child in every school to which grants were to be paid (Royal Commission, 1861). This system—which perhaps predictably came to be known as "payment by results"—has been the subject of much analysis and debate. Some (e.g., Hurt, 1971 p. 222) have argued that the focus on reading and arithmetic was essential to breaking the monopoly of the religious schools, and furthermore, that making schools and teachers accountable through the use of objective measures of achievement was necessary to demonstrate to the wider public the political case for investment in publicly-funded elementary education (e.g., Hurt, 1971; Sylvester 1974; Mitch, 1999). Others (e.g., Rapple, 1994) have countered that the damage done by the high-stakes accountability regime was worse than the problem it was trying to address:

> true accountability in education should not be facilely linked to mechanical examination results, for there is a very distinct danger that the pedagogical methods employed to attain those results will themselves be mechanical and the education of children will be so much the worse. (p. 21)

In the United States, at the same time, there were many who were concerned about the damage done, both to students and teachers, by the introduction of "high-stakes" assessments. In his *Elements of Pedagogy*, Emerson E. White considered "the propriety of making the results of written examinations the basis for the bestowment of scholastic rewards and honors, for the promotion and classification of pupils and for determining the comparative standing or success of schools and teachers." (White, 1886 p. 198) His conclusion was forthright:

> They have perverted the best efforts of teachers, and narrowed and grooved their instruction; they have occasioned and made well-nigh imperative the use of mechanical and rote methods of teaching; they have occasioned cramming and the most vicious habits of study; they have caused much of the overpressure charged upon schools, some of which is real; they have tempted both teachers and pupils to dishonesty; and last but not least, they have permitted a mechanical method of school supervision. (pp. 199-200)

Haertel and Herman (2005) provide a brief overview of the last century of accountability testing in the United States. They conclude:

> From the days of Joseph Rice and the school testing programs of the early 1900s, through the Head Start program evaluations of the 1960s, and up to the increasingly prescriptive testing requirements of successive ESEA [Elementary and Secondary Education Act] reauthorizations culminating in NCLB [the 2001 No Child Left Behind Act] policymakers have used tests in an attempt to discover which schools and districts are fulfilling their responsibilities and which are falling short. (pp. 28-29)

One of the distinctive features of these approaches to "testing for accountability" in the United States is that the stakes were much higher for teachers than for students. Indeed, apart from any actions that teachers might have taken against students who performed badly on these tests, there were no repercussions at all for the students. Tests that are high stakes for teachers but low stakes for students have been widely used in the United States for many years, but are relatively rare in other developed and developing countries, where the testing regimes are either low stakes for teachers and high stakes for students or high-stakes for both. The reasons for the divergence in practice between the United States on the one hand and what I shall call here the European tradition in testing (which for these purposes includes Japan) are complex, and beyond the scope of this article. However, one factor appears to be particularly important, and that is the difference in view of the purpose of schooling, particularly for adolescents.

Upper secondary schooling: divergent aspirations

Within the European tradition, at least up until the last quarter of the twentieth century, education beyond the age of 15 or 16 was intended only for the small proportion of the national population planning to enter higher education (typically only five or ten percent). As a result, the assessment arrangements in place at the end of high school were determined by the universities, and provided a *de facto* curriculum for upper secondary schools. Indeed, what is called the "syllabus" in secondary schools in England is simply a list of all the things that can be tested in the examination. Alignment between curriculum and instruction was total, since the only purpose of being in school beyond the statutory leaving age was in order to pursue entry to a university, and therefore for most students, the curriculum consisted entirely of test preparation.

The assessment was also standards-based in that the assessment was largely, if not entirely, based on achievement on timed, constructed-response, achievement tests based on carefully prescribed syllabuses. Secondary education systems in Europe have the appearance of being designed backwards from the point of selection into higher education with a series of "hurdles" for the student to clear (e.g., selection into a *Gymnasium* at the age of 10 in Germany; selection into a selective "grammar school" in England at the age of 11; selection into one of the most prestigious middle schools in Japan).

In contrast, in the United States, there appears to have been a widespread consensus about the value of education up to the age of 18 as a general preparation for

adulthood, complete in and of itself, which led to the extraordinary expansion of secondary education in the United States between 1910 and 1940 (Goldin, 2002 p. 25). Because high school was intended for *all* students, it would have been inappropriate to assess students against standards intended for the small proportion going on to higher education. Moreover, within the United States, the process of *selection* to a higher education institution and *placement* in a particular program were separate, so that it was possible to select on the basis of general aptitude, delaying the placement decision for a year or two. In most European universities, the processes of selection and placement are combined—one cannot apply for a place at a particular university without specifying a particular program—so that universities need to establish an applicant's aptitude for a particular program before deciding whether to admit the applicant or not.

There were some states, such as New York, that imposed state-wide university entrance examinations, but for the rest of the United States, high-school graduation requirements were generally determined by districts. In the 1970s, a number of states introduced minimum competency tests for graduation (see Madaus, 1983, for a summary), but as Pipho (2002) points out, even here, the extent to which these tests have been high-stakes for the students has varied considerably, and in many districts, schools continued with curricula that were only weakly aligned with the standardized tests routinely used to hold schools accountable. Why the use of such standardized multiple-choice tests persisted in the United States, despite the poor alignment with what schools were teaching, when such tests are largely anathema in other rich countries (Wood, 1991) is obviously not a simple question, and probably involves

matters of cultural preference as much as technical matters such as reliability and other aspects of validity (Black & Wiliam, 2005). Whatever the reason, the consequences of this lack of alignment are discussed in the section entitled "Why are relative school effects so small?" below.

The logic of accountability testing

The logic of accountability testing is deceptively simple. Students attending higher-quality schools will (by definition) have higher achievement than those attending lower-quality schools, so that differences in the quality of schooling will result in systematic differences in achievement between schools. Provided the accountability tests assess school achievement, then higher test scores will indicate higher-quality schooling. However, what is required for school accountability is the converse: that higher student scores are indicative of higher-quality schooling.

As Messick (1989) has pointed out, this is a matter of validity, which he defined as "an integrative evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment." (p. 13) In Messick's formulation, the two main threats to valid inference based on test scores are construct under-representation and construct-irrelevant variance. The former occurs when test scores fail to represent adequately differences in student achievement on the construct of interest, and the latter occurs when differences in student achievement are not related to the construct of interest (put simply, construct under-

representation indicates an assessment that is too limited, in that it fails to assess things it should, while construct-irrelevant variance indicates an assessment that is too large, in that it assesses things that it should not). In the specific context of accountability testing, construct under-representation is minimized by ensuring that, as far as possible, the tests assess what schools are intended to develop in their

students, while construct-irrelevant variance is minimized by seeking to ensure

that differences in student scores are related to differences in the quality of schooling received, rather than differences in the students' aptitude, socioeconomic status and so on.

The rise and fall of authentic assessment in the United States

It was a concern with a form of construct under-representation—that standardized tests (and particularly many of the "minimum competency" introduced in the 1970s) were not assessing all important aspects of school achievement—that led, in the 1980s and 1990s, to increased interest in the assessment of so-called "authentic work," often through portfolios of student work. Although there is evidence that such approaches can have significant positive impact on student learning (Newmann, Bryk & Nagaoka, 2001), it is also clear that such approaches fell far short of traditional standardized tests in psychometric terms—especially in respect of reliability (Koretz, Stecher, Klein, McCaffrey & Deibert, 1994). This, combined with the high cost of such assessment, led most states either to completely discontinue the use of constructed response assessments, or to scale back their use substantially.

On January 8th, 2002, President George W. Bush signed into law the 2001 No Child Left Behind (NCLB) Act (technically a re-authorization of the Elementary and Secondary Education Act, first signed into law in April 1965, with major changes introduced through reauthorizations in 1967, 1981 and 1994). Although unprecedented in its powers, and in the detail in which its provisions were laid out, NCLB can be seen as an evolution of previous attempts to use high-stakes tests to improve educational outcomes. Like its predecessors, the stakes are much greater for teachers, schools and districts than they are for individual students[1]. It focuses on the performance of successive cohorts of students, rather than individual students, so that in each school, what matters is that the proportion of each student in each grade (from third to eighth grade) achieving proficiency (however it is defined by the state) increases each year, and reaches 100% by 2014.

At the heart of NCLB is the logic of accountability testing outlined above: differences between students in terms of their educational outcomes, as measured by the tests, should be largely, if not wholly, attributable to differences in the quality of education provided by schools. Whether such inferences are valid depend on the extent to which the tests used adequately represent the construct of interest (what schools are designed to develop in their students) and the extent to which differences in test scores represent differences in the quality of schooling, rather than other factors.

Whether accountability tests adequately represent the construct of interest is a complex matter, requiring a clear definition of the construct of interest (see Wiliam, 2010), which is beyond the scope of this article. The remainder of this article focuses

on the issue of construct-irrelevant variance; to what extent do differences in test scores represent differences in the quality of schooling provided (construct relevant variance) rather than other factors, such as the amount of parental support, differences in the prior achievement of students on entering the school system, and so on (construct-irrelevant variance)?

What are the sources of variability in test scores?

For over a hundred years, researchers have explored sources of variability in test scores (Spearman, 1904; Mackintosh, 2000; Deary, Strand, Smith & Fernandes, 2007). Some of these studies have tried to disentangle the effects of some measure of general intelligence from socioeconomic status, but when tests are used to hold schools accountable, such distinctions are much less important than how much of the variation in test scores is between schools and how much is within schools. This is because if differences in test scores are due to differences in school quality, then the proportion of variance that is between schools provides an upper bound on the proportion of variance that is attributable to school quality. This is an upper bound, because systematic differences in, for example, levels of school funding will also contribute to between-school variation, but is not a matter that the individual school can influence (although in most states, the district can).

As part of its Programme for International Student Assessment (PISA), the Organisation for Economic Cooperation and Development (OECD) has developed a range of measures of student knowledge and skills, specifically focusing on reading,

mathematical, and scientific literacy (OECD, 2000). The development of the framework for assessment, and of the measures, is summarized in McGaw (2008). The purpose of PISA was to answer broad questions about the quality of education systems, such as:

> How well are young adults prepared to meet the challenges of the future? Are they able to analyse, reason and communicate their ideas effectively? Do they have the capacity to continue learning throughout life? Parents, students the public and those who run education systems need to know the answers to these questions (OECD, 2000, p. 3).

In order to address these questions, an explicit decision was taken within PISA not to attempt to assess the extent to which student had mastered knowledge and skills specified in their curricula that shaped their schooling, as, for example, had been a clear focus within the Trends in Mathematics and Science Study (TIMSS) program (Mullis, Martin, Ruddock, O'Sullivan, Arora, & Erberber, 2005). Instead, PISA focused on the capacity of a specific population of students—those reaching the end of compulsory schooling— to reflect on and use the skills they have developed (McGaw, 2008).

One particularly significant strand of analysis within PISA was to apportion the variance in student scores on the PISA tests into between-school and within-school variance components. In the 2003 mathematics assessment, variance in student achievement in the United States was about 5% larger than the average of the

participating countries, approximately one fourth (25.7%) of the variance being between schools, and three-fourths of the variance was within schools. It might be assumed that this between-school variance represents the "school effect"—the difference in student achievement that is attributable to the school, but in the United States, students are not randomly allocated to schools, so that the between-school variance includes both the "school effect" and systematic differences in the students attending different schools (in the United States, more affluent students are clustered together, as are less affluent students). In order to investigate the extent to which between-school differences were caused by differences in the composition of the students attending the school, rather than differences in school quality, PISA developed an international index of economic, social and cultural status. Using this index, it was found that more than two-thirds (69%) of the between-school variance was accounted for by differences in the social class of the individual students and by further contributions of the average social class of other students attending the school (sometimes called "compositional" effects). This means that no more than 8% (31% of 25.7%) of the total variance in mathematics achievement of 15-year-olds in the USA was attributable to the quality of the education provided by the school (OECD, 2004). The results in science are similar (OECD, 2007).

Now of course this does not mean that schooling does not make a difference. In the countries surveyed in PISA schooling up to the age of 15 is practically universal, so we are not comparing students who have, and have not been educated. We are therefore unable to draw any inferences from these data about the effects of schooling. However, it is reasonable to draw inferences about the extent to which *variations* in

student achievement is associated with *variations* in the quality of schooling, which is one of the main purposes stated by advocates of high-stakes accountability testing—the idea is that differences in test scores are indicative of differences in the quality of education provided.

Whether measures of student achievement for purposes of accountability should control for compositional effects—and specifically the extent to which high-achieving students increase the achievement of other students alongside whom they are taught—is not straightforward. As has been recognized for some time (Wiliam, 1992; Willms, 1992) different audiences for school accountability data will require different kinds of analyses. Parents choosing schools for their children may well not want measures of school performance to control for compositional effects—put bluntly they may not care whether their children do well because they are well taught or because there are other able children at the school who increase the achievement of others around them. On the other hand, it seems unfair to reward some schools and sanction others simply for their success (or lack of it) in attracting high-achieving students.

Value-added measures of school effectiveness

In England, over the past ten years, there has been a substantial investment of effort in the development of ways of disaggregating some of the factors contributing to the progress made by students at schools. This work has been greatly facilitated by the "vertical design" of the national curriculum in England, which focuses on learning

progressions over time, rather than allocating different aspects of content to different grades (Wiliam, 2007).

Achievement of 16-year-olds in England is measured primarily through a set of national examinations entitled the General Certificate of Secondary Education, or GCSE, although many vocational programmes are also available. Most students sit examinations in 12 to 14 subjects at GCSE, and their results are reported on a nine-point scale (U, G, F, E, D, C, B, A, A*), the grades being derived by setting cut-scores on a continuous scale involving a weighted combination of scores derived from externally set examination papers and school-based assessment (the balance between the two differs for different subjects). The most common measure of school performance for secondary schools in England—and the one that is featured in national daily newspapers when the results are published—is the proportion of the age-16 cohort at the school achieving one of the four highest grades (C, B, A, A*) in English, mathematics and at least three other subjects in the GCSE, often abbreviated to "5 A* to C including English and mathematics." This proportion varies from close to zero in schools catering predominantly for students with profound special needs up to 100% for highly selective schools.

For many years there has been a concern that such an index is difficult to interpret. Do high scores indicate high-quality schooling, or simply a school that attracts high-achieving students? To address these concerns, a number of approaches to distinguishing the contributions of the school from other factors, such as the prior attainment of the students, have been explored.

In the United States, such "value-added" approaches have, in general, been more directed towards the identification of the effects of individual teachers on student achievement, often designed to support differentiated compensation (see, for example, Braun, 2005) although within the last two years, pilots of value-added approaches to determining whether a school has made "adequate yearly progress" towards the goals established by the state under the No Child Left Behind Act have been authorized. In contrast, in England, most approaches to value-added have focused on school-level, rather than teacher-level effects.

Value-added analyses of school performance have been greatly assisted by the fact that the national curriculum for England was designed in terms of "learning progressions" (Wiliam, 2007; Heritage, 2008), and a number of relatively simple models for estimating the contribution of the school to a student's achievement have been developed (Wiliam, 1992, Jesson & Crossley, 2007).

The earliest models used for reporting on the value-added by secondary schools converted the best eight grades achieved by a student at GCSE (or the equivalent vocational programmes) into a numerical score (0, 16, 22, 28, 34, 40, 46, 52 and 58 respectively for the nine grades) and regressed the total score on a similar measure based on the scores achieved by the student in the national tests at age 11. For each student, a residual was calculated, and the relative "value-added" by the school was simply the average of the residuals for all students in the age 16 cohort (for further details of the procedure, see Ray, 2006 and Kent & Blows, 2009). Analysis of these

models revealed that a significant proportion of the variance attributed to the school was due to contextual factors (e.g., girls had larger residuals than boys, so that the higher value-added estimates for single-sex girls' schools were in part due to the differences in intake rather than differences in school quality) and so a number of refinements have been made to the original model to produce an estimate of the so-called "contextualized value added" provided by the school.

In addition to prior attainment, the model currently in use includes adjustments to the predictions for each student to take into account sex, ethnicity, poverty (as measured by entitlement to free school meals, and an area-based measure of the proportion of households on low incomes), the extent of any special educational needs, English language proficiency and a number of other smaller adjustments (Ray, 2006). It was also discovered that the relationship between achievement at age 11 and achievement at age 16 was not linear but curved slightly upwards (when achievement at 11 is plotted on the x-axis and achievement at 16 is plotted on the y-axis). The CVA model also therefore includes a quadratic term, so that the achievement at age 11 and the square of the achievement at age 11 are fed into the model to improve the fit of the model. The resulting measure of contextualized value-added (CVA) is reported on a scale anchored to the grades of the GCSE, so that the score for a school is reasonably easily interpretable[2].

For the 4158 schools in England that had students taking the GCSE in 2007, the correlation between the most common outcome measure—the proportion of students in a cohort achieving at least a grade C in 5 subjects in the GCSE—and the CVA

measure was 0.27, indicating that the school effect contributes only around 7% of the variance in student outcomes: quite close to the 8% estimate generated by PISA for the United States. Put another way, one standard deviation of school quality (as measured by CVA) equates to about one-fifth of a standard deviation of student achievement. Since the proportion of students at a school gaining five good grades at GCSE including English and Mathematics is based on a discrete measure (each student at a school either does, or does not, achieve this), the correlation between this and the CVA will be attenuated to some extent, but since the vast majority of schools have at least 150 students in an age cohort, the reduction in correlation is likely to be small, if not negligible. Differences in school quality, both in the United Kingdom and in the United States, would therefore appear to contribute relatively little to the differences in student achievement.

The remainder of this article explores some possible reasons why relative school effects contribute so little to student outcomes, and what happens when standardized tests are used to hold schools accountable. The article does not deal, therefore, with broader notions of accountability (see, for example, Darling-Hammond, 2006) nor does it deal with the idea that accountability should be a "two-way street" in which schools are accountable to the community for student outcomes, but also where the community is accountable to the school for the provision of adequate resources for the task (Chambers, Parrish, Levin, Smith, Guthrie, Seder & Taylor, 2004) through so-called "adequacy lawsuits" (Hanushek, 2006). These are important issues, but beyond the scope of the present article.

Why are relative school effects small?

There are many reasons why between-school effects account for a relatively small proportion of the variance in student test scores. Any student score will involve some error, and this will reduce the between-school variance. In addition, no measure will perfectly capture the kinds of changes that schools effect in their students. Where the alignment between the test and the curriculum is particularly poor, scores on the test will substantially under-represent the construct of interest, namely mastery of the curriculum, so that the scores of students experiencing higher-quality schooling (defined as schooling producing better mastery of the specified curriculum) will not necessarily be higher than those experiencing lower-quality schooling. In this section, I explore in detail a further reason, namely that the rate of progress of individual learners is small in comparison with the range of achievement within a cohort. Specifically, I show that the range of achievement within an age cohort can be ten or fifteen times greater than the amount learned by a typical individual in a year. The result is that even large increases in the rate at which individual students learn will produce small increases in the proportion of variance in student outcome measures attributable to the school.

The Leverhulme Numeracy Research Programme (LNRP) administered a series of numeracy tests to two cohorts of elementary school students in England over a four-year period. One cohort began in kindergarten, and the other began in third grade, and each participating student was tested twice each year (in October and May or June). In order to make the tests appropriate for students of different ages, the tests varied from grade to grade, but eleven items were used across five grades, allowing the

increase in facility for a particular item to be tracked from kindergarten to fifth grade. One item (code 1106), presented orally to students in first through fifth grade, asked students to complete the following calculation:

"Eight hundred and sixty add five hundred and seventy"[1]

Although the item was presented orally, students were provided with scratch paper for calculation, and were asked to write their answer in a space given to them in the test booklet. The facility of the item for students of different ages is as shown graphically in Figure 1.

Figure 1 about here

One interesting feature of Figure 1 is how slowly the facility of this item increases with age. Approximately twenty percent of students can answer the item correctly by the age of 8, but three years later, thirty percent still cannot. The testing of the students was undertaken as part of a research program within which teachers and students were guaranteed anonymity. The test was therefore low-stakes for both teachers and students. Since this was a nationally representative sample, and since the timing of the test was not keyed to any curricular coverage of the specific skills being tested, it seems reasonable to assume that this is a fairly accurate indication of the "response to treatment" under conditions of typical instruction. One explanation for

[1] This is the correct form of expression for these numbers in British English.

the relatively small increase in facility over time could be that this was not a skill that teachers taught. However, in a teacher questionnaire, all teachers participating in the program were asked to indicate, for each item, whether the skills being assessed were skills they sought to develop in students, and almost all teachers (well over 90%) indicated that the skill tested by this item was something they taught and reviewed regularly with students. Furthermore, since the sharpest increases in facility occur between October and June, rather than between June and October, it seems likely that attendance at school, rather than general maturation, is the main cause of the increase in facility.

Of course this is just a single isolated item, but across the 159 items used in the LNRP tests across the six grades, the average annual increase in facility was just sixteen percentage points. Each of these items was regarded as grade-appropriate for the grade in which it was tested, and the teachers agreed that the item assessed a skill that they were trying to develop in their students that year. For the easier items, there will be "ceiling effects"—once the facility exceeds 84%, then of course increasing the facility by sixteen percentage points is impossible. However, even if we exclude the items with facility greater than 85%, then the average annual increase in facility is only twenty-five percentage points. In a class of 24 students, an annual increase of item facility of twenty-five percentage points suggests that only six students are acquiring the skill in any given year. The other 18 students in the class would either already know it at the beginning of the year, or still would not know it at the end of the year.

The fundamental idea here—that the rate of progress of individual students is slow compared to the range of achievement within an age cohort—is not new. Two reports from the Assessment of Performance Unit (APU) in the United Kingdom (roughly similar in purpose to the National Assessment of Educational Progress or NAEP in the USA) in 1980 had shown that high-achieving 7-year-old students out-performed some 14-year-olds on basic arithmetic. One item in particular:

$$6099 + 1 = ?$$

gained some notoriety when it was found that there were many 14-year-olds who thought the answer was 7000, while many 7-year-olds knew the answer to be 6100 (Foxman, Cresswell, Ward, Badger, Tuson & Bloomfield, 1980; Foxman, Martini, Tuson & Cresswell, 1980). One influential official inquiry branded this phenomenon as a "seven-year-gap" between the lowest and highest achieving students in a middle-school mathematics class (Committee of Inquiry into the Teaching of Mathematics in Schools, 1982).

Other studies of mathematical abilities in school students showed the same, or even greater, variability than had been found by the APU. The Concepts in Secondary Mathematics and Science (CSMS) project had identified a series of six age-independent levels of understanding of decimals, and, in a nationally representative sample, found that the variability within each age cohort was much greater than the differences between cohorts (Hart, 1981). In particular, in a cross-sectional study of achievement, involving at least 500 students at each age point, the proportion of

students achieving a particular level increased by no more than 10% per year (see Figure 2). These studies were recently replicated, and while there were some noticeable changes (for example, success rates on items assessing decimals had improved, while scores on items assessing fractions had deteriorated), the overall variability of achievement had changed little (Hodgen, Küchemann, Brown & Coe, 2009).

Figure 2 about here

Brown, Blondel, Simon and Black (1995) found that in England, performance on conceptual issues involved in measuring length and weight (mass) were also slow to develop. There were some first-grade students who were well ahead of *most* students in seventh grade, suggesting that there may be as much as a "twelve-year gap" between the weakest and the strongest in a seventh grade science class, which is consistent with similarly-focused research in mathematics (Brown, 1992 p. 12).

Such findings are also typical in the United States. In the mid-1950s the Cooperative Test Division at the Educational Testing Service produced a series of Sequential Tests of Educational Progress (STEP) in reading, writing, listening, social studies, mathematics and science (Educational Testing Service Cooperative Test Division, 1957). The tests were aimed at students from 5th grade to the first two years of college, and were vertically scaled, permitting comparisons to be made across years. The annual increase in achievement in the STEP tests, measured in standard

deviations, is shown in Figure 3. Apart from the earliest and latest grades, the typical annual increase in achievement is between 0.3 and 0.4 standard deviations.

Figure 3 about here

Other tests show similar properties. Petersen, Kolen and Hoover (1989) discuss the results of scaling the results from the Iowa test of basic skills (ITBS) language usage test (Hieronymous & Lindquist, 1974) for different cohorts of students. By definition, a median grade 3 student attains a grade equivalent of 3.5 half-way through the year. The data from the ITBS scaling studies indicate that about 30% of students will, by half way through grade 3, have achieved a score equivalent to that achieved by the median fourth-grader at the same time. In a very real sense, therefore, these 30% of students are at least one year ahead of the median student in their grade. Collecting similar data points for third graders and joining them up generates a "grade characteristic curve" for third grade. A similar analysis applied to students in other grades produces a series of such curves (see Petersen, Kolen & Hoover, 1989 p. 234). So, for example, in the ITBS language usage tests, the standard associated with average students half way through fourth grade is also just attained by the lowest attaining 5% of students in eighth grade, the lowest-attaining 10% of those in seventh grade, the lowest-attaining 18% of those in sixth grade, and the lowest-attaining 30% of those in fifth grade. On the other hand, the same standard is reached by the highest-attaining 30% in third grade as noted above, and probably by some students in second grade, although this is not recorded.

One response to this line of argument is that such grade-to-grade comparisons are difficult to interpret, since strictly speaking, a grade-equivalent score has meaning only within that grade (as Petersen *et al.* point out). However, in domains such as reading and mathematics, it would seem rather odd to claim that a curriculum for one grade does not build on that for the previous grade, so that there is a sense that, even in language usage, some third-graders are performing like some eighth-graders and vice-versa. In the ITBS test, one year's growth ranges from around 0.5 standard deviations in third-grade, to around 0.35 standard deviations in 8th grade, which is quite similar to the data for the STEP tests shown in Figure 3.

More recent data has confirmed that one year's growth in achievement typically ranges from around 0.25 to 0.4 standard deviations. Rodriguez (2004) found that one year's progress in middle-school mathematics on the tests used in TIMSS (Trends in Mathematics and Science Study) was equivalent to 0.36 standard deviations, while the average increase in achievement in mathematics from fourth-grade to eighth-grade on the assessments used in the National Assessment of Educational Progress (NAEP) is approximately one standard deviation (NAEP, 2006), suggesting that for the NAEP tests, one year's growth is only about one-fourth of a standard deviation.

Most recently, Wibowo, Hendrawan and Deville (2009) reported their attempts to design a vertical scale for the reporting of student achievement on accountability tests in Connecticut. While growth in fourth grade averaged over one standard deviation in mathematics, over the subsequent four years, the average growth in achievement in mathematics and reading was less than a third of a standard deviation (Figure 4).

Figure 4 about here

Moreover, these effects are not just limited to standardized tests in K-12 education. Norcini (2009) reports that in cardiac surgery, the effects of one year's training is approximately equivalent to one-third of a standard deviation.

The fact that the average effects of one year's instruction, on a range of outcome measures including standardized tests, is often as little as a fourth of a standard deviation, typically around one third of a standard deviation, and rarely more than a half of a standard deviation, suggests that test scores are measuring much more than just the quality of instruction. After all, if these tests were measuring *only* the effects of instruction, then a naïve analysis would suggest that the gap between (say) the average achievement of cohorts of seventh grade and eighth grade students could be expected to be as much as four standard deviations on the grounds that those eighth graders receiving the least effective instruction should be performing no worse than those receiving no eighth grade instruction (i.e., seventh graders).

To sum up so far, because the progress of individual students is slow compared to the variability of achievement within the age cohort, variance in students' scores is much more strongly related to features over which schools have little influence, such as the prior achievement of students, than to the quality of the education provided by the school, and therefore are not well-suited to supporting inferences about the quality of

the education provided by a school. The next section deals with the impact of such tests on student outcomes.

The impact of high-stakes accountability testing on student achievement

An evaluation of the impact of NCLB is beyond the scope of this article, and in any case, would at this stage be premature. However, as the Obama administration considers the re-authorization of NCLB, it is useful to review briefly the existing research on the impact of high-stakes accountability testing on student achievement.

It has been widely observed for many years that when any test (or indeed any other performance indicator) is made the focus of public policy attention, then performance as measured by that test improves—an effect known as Campbell's Law in the United States (Campbell, 1976 p. 49) and Goodhart's Law in the United Kingdom (Kellner, 1997). Perhaps the best known example of this is the case documented in Koretz, Linn, Dunbar and Shepard (1991) in which it was found that a district's scores on a newly introduced test started low, but improved steadily, while scores on the test whose use had been discontinued declined. In fact, Cannell (1988) found that all fifty states posted state averages above the national average, a situation that was described as the Lake Wobegon effect (after Garrison Keillor's mythical town where all the women were strong, all the men were good looking, and all the children were above average). There is little doubt, therefore, that attaching high stakes to test outcomes can increase the scores on those tests. The important question, however, is whether such improvements generalize to other tests that are intended to measure

the same constructs, and in particular whether they generalize to *remote* and *distal* measures (Ruiz-Primo, Shavelson, Hamilton, Klein, 2002).

In a widely reported analysis, Amrein and Berliner (2002a) examined the impact of the introduction of high-stakes testing programs in 18 states. They concluded that while there was clear evidence that associating high-stakes consequences to test score outcomes had increased scores on the tests used within the program, there was no evidence of improved test scores on other related measures, such as the College Board's SAT and Advanced Placement tests, the ACT (formerly American College Testing) test and on the National Assessment of Educational Progress (NAEP). Furthermore, they found that the introduction of high-stakes testing regimes was associated, in some cases, with increased student drop-out rates, inappropriate test preparation practices (up to and including cheating), and decreased teacher morale, leading to increased teacher defection from the profession. A subsequent analysis, involving the 27 states with the highest stakes associated with test score outcomes in grades 1 through 8 (Amrein and Berliner, 2002b) confirmed these findings, and also indicated that the introduction of high-school graduation examinations was associated with a lowering of average academic achievement.

However, as Rosenshine (2003) pointed out, when results for the states with the "clearest" high-stakes policies were compared with those without high stakes, then over the most recent four-year period for which data were available, high-stakes testing regimes were associated with greater increases in NAEP scores. Moreover these effects were quite large (standardized effect size of 0.35 for 4th grade

mathematics, 0.79 for 8th grade mathematics, and 0.61 for 4th grade reading), especially in view of the relative insensitivity of NAEP tests to the effects of instruction as noted above, although the use of NAEP as the "anchor" for all these studies is a significant limitation.

Hanushek and Raymond (2005) also re-examined the data used by Amrein and Berliner using more sophisticated methods, and found that in the states with the strongest accountability regimes, the increase in scores on NAEP from fourth grade in 1996 to eighth grade in 2000 was three points (around 0.2 standard deviations) greater than in those with the weakest accountability regimes, holding other inputs and policies constant. They also found that "report cards" on school effectiveness were not a significant factor, suggesting that it was the direct effect of the incentives rather than the information provided about school performance that was producing the increase in achievement. However, while the net effect of accountability regimes was to increase student achievement, the effects were different for different minorities, with Hispanics gaining most, and African-Americans gaining least, so that the introduction of accountability regimes appears to widen, rather than narrow the achievement gap between white and African-American students.

Rather than simply dichotomizing states as having either low-stakes or high stakes testing regimes, Carnoy and Loeb (2002) developed an index of the strength of the accountability system in place in each of the 50 United States that assigned a score between 0 (low) and 5 (high) according to a range of factors such as the number of grades in which testing was mandated, the repercussions for schools, the presence of

high school exit examinations and the length of time the accountability regime had been in place. Using the accountability index as a continuous variable in regression analyses, they found that stronger accountability regimes were associated with greater increases in NAEP scores at eighth grade (although not at fourth grade, which they considered surprising in view of the fact that gains in state test scores tend to be larger at fourth grade than eighth grade). For example, a two-step increase in the accountability index (e.g., from 1 to 3) was associated with an increase of 2.8 percentage points in the proportion of white eighth-grade students classed as having reached the "basic" level on NAEP, and this was more than half the national increase over the same time period (1996 to 2000). In contrast to Raymond and Hanushek, Carnoy and Loeb found that high-stakes accountability regimes benefitted all minorities; for African-American students, a two step increase in the accountability index was associated with an increase of 5.1 percentage points (national average increase 5.7), and, for Hispanic students, larger still (8.9 percentage points compared to a national average increase of 6.1). Also in contradiction of the findings of Amrein and Berliner, they found no evidence of increased rates of retention, nor of lower rates of high school completion, although, perhaps surprisingly, increased achievement in mathematics achievement at eighth grade did not seem to lead to increased progression through high school.

Braun (2004) undertook an extensive re-analysis of the performance of states from 1992 to 2000 on NAEP at fourth and eighth grade. Using a variety of sophisticated models, he confirmed the earlier findings of Hanushek and Raymond (2005) and Carnoy and Loeb (2002), namely that high-stakes accountability regimes were

associated with greater increases in NAEP scores at eighth grade (but not at fourth grade). However, perhaps more importantly, he showed that the association between NAEP score gains and high-stakes testing regimes disappeared in cohort-based analyses. Reviewing 14 studies on the impact of accountability regimes (and 76 effect sizes derived therefrom), Lee (2008) showed that the average effect sizes for grade-based (i.e., cross-sectional) analyses was 0.40, but was only 0.03 for cohort-based analyses. In addition, he pointed out that in describing effect sizes, it is important to be clear about whether the divisor in the calculation is the standard deviation of the state-level gain scores or the student-level scores. Recalculating all 76 effect size estimates in the 14 studies using student-level, as opposed to state-level, standard deviations produces a mean effect size across all the studies of just 0.08 standard deviations. While this might be regarded as a small effect when judged in terms of generally accepted interpretations of standardized effect sizes (Cohen, 1988), given that one year's average growth is typically around 0.4 standard deviations, this effect equates to a 20% increase in educational productivity, which would have a substantial cash value if replicated across the entire United States.

One of the problems with the research reviewed above is that the testing regimes were classified only in terms of the extent to which high stakes were attached to the outcomes of the test. Analyses by Bishop (2001a, b) showed that the nature of the tests is also important. Specifically, he found that students who were required to pass externally-set curriculum-based examinations—such as those found in New York and North Carolina, and indeed in most developed countries other than the United States—learned more than those who did not.

To summarize this section, although there is a need for more research, particularly into the impact of different kinds of tests, there is growing evidence that the impact of high-stakes assessments are not simply confined to performance on the tests themselves. The presence of high-stakes assessment systems appears to increase student achievement on a range of measures that are distal, or even remote, from the accountability tests. In most of the research cited above, the "reference instrument" has been NAEP tests, which were designed to be quite general measures of achievement, applicable in all states, despite the differences in curricula. The fact that the introduction of a high-stakes accountability system increases scores on NAEP as well as on state-mandated tests indicates that the effects of accountability tests generalize well beyond what is actually tested (Phelps, 2005). The immediate question, therefore, is whether there are additional costs that would outweigh the benefits.

Unintended outcomes of accountability testing

As the research on the impact of accountability testing has shown more and more clearly that, under a wide range of conditions, accountability testing can raise student achievement on a broad range of measures, much of the critique has focused on the "collateral damage"—the unintended consequences of the use of such accountability regimes.

Cizek (2005) identifies a range of unintended outcomes that have been claimed in critiques of accountability tests including a disproportionate focus on tested content, demoralization of teachers, and undue pressure on students. However, after identifying one study in two elementary schools that identified a range of issues that needed further exploration, he suggests that few, if any, of these studies provided strong support for the claim of adverse impact on student achievement: "In the case of high-stakes testing critiques […] the subsequent evidence collection appears to have become even more skimpy in support of conclusions that seem even more confident" (p. 28).

Harlen and Deakin-Crick (2002) undertook a systematic review of the impact of tests and other standardized forms of assessment on students' motivation for learning. An initial search identified 183 potentially relevant studies of which 19 were found to be directly relevant to the focus of the review. Of the 19 studies, only 13 explicitly examined student outcomes (three randomized controlled trials, three case control designs, three post-test and four others). Although the authors claim strong evidence for a number of findings also found in other critiques (teachers emphasizing lower-order skills, lowered self-image and increased anxiety for lower-achieving students, a shift from mastery-orientation to performance orientation and extrinsic motivation) most of the evidence for these negative effects was found in naturalistic studies that provided insufficient details of the context of data collection to generalize the findings. At this stage, therefore, the evidence about the negative effects of high-stakes testing would appear to be inconclusive.

There is evidence that high-stakes accountability testing makes it harder to keep teachers (Clotfelter, Ladd, Vigdor & Diaz, 2003), that teachers of disadvantaged students are likely to experience greater pressure to improve their test scores and to focus on test content than teachers of more advantaged students (Herman, Abedi & Golan, 1994), as well as a host of other unintended outcomes. However, given the evidence that accountability systems can have positive effects on student achievement, it would seem appropriate to explore whether the negative effects of high stakes assessment might be ameliorated, while still maintaining the positive impact described above.

Improving high-stakes accountability tests

There are a number of measures that could be taken to ameliorate some of the difficulties with high-stakes accountability testing identified above. However, until such measures are implemented and evaluated, any proposals would be at best tentative. Having said this, perhaps the most significant, and best developed, proposals in this regard have come from a group commissioned by the National Research Council (and funded by the National Science Foundation) to investigate the development of "instructionally-supportive accountability tests" (Popham, , Keller, Moulding, Pellegrino, & Sandifer, 2005). In its final report, the group made a number of recommendations for the improvement of high-stakes accountability testing in science, although the recommendations would appear to be generalizable to any school subject.

The group proposed a number of conditions that would need to be met if accountability tests were to be instructionally supportive. The first was a reduction in the number of curricular aims. They noted that in many states there can be more than fifty content standards for a subject at each grade, each of which can contain multiple aspects. They suggested that curricular aims should be defined more broadly around key concepts or ideas and ideally there should be only around a dozen or so such aims for each grade. The distillation of current content standards into such a small number of aims is clearly a far from trivial task. One possibility, discussed in some detail in the group's report, involves the use of a "skill-by-concept" matrix.

A group of science educators had agreed that in their teaching of physical science in middle school, there were four main concepts: characteristic physical properties and changes, characteristic chemical properties and changes, forces and motion, and forms of energy and energy transfer. In addition, they identified five "science-as-inquiry" skills (posing questions; designing investigations; gathering, analyzing and interpreting data; developing descriptions, explanations, predictions, and models; and thinking critically about links between evidence and explanation). Crossing the four concepts with the five skills generated a skill-by-concept matrix with 20 cells, a dozen of which might be selected for a particular grade. While the group acknowledges that a different group of educators would almost certainly generate a different skill-by-concept matrix, the *process* identified above does appear to be a robust and general method "for deriving a modest number of high-import curricular foci from a state's existing curricular aims in science." (p. 138)

The second condition was that curricular aims should be expressed in language that is comprehensible to teachers. The group observed that where teachers do not understand the meaning of the curricular aims as expressed in the state's content standards, they are likely to key their instruction on sets of test items relating to that standard, rather than to the aim that the items are intended to represent. They suggested that the assessment descriptions that that accompanied a content standard should be brief, written in straightforward language, and should include illustrative test items.

The third condition was that the tests should be designed, and their results reported, in such a way that *each* student's mastery of *each* curricular aim can be assessed with reasonable accuracy. Traditional considerations of validity (including reliability) would indicate that the tests should contain a number of items on each curricular aim. However, even a dozen or so curricular aims would be too many to assess in the limited amount of time available for statewide testing. Rather than increase the amount of time taken by testing, the group suggested that in any one year, only a few aims (i.e., three to five) would be assessed. To avoid the dangers of what they termed "curricular reductionism" (teachers teaching only those aims that were going to be tested), the group recommended that for a given year's accountability test, all of the dozen or so curricular aims would be eligible for assessment, but the number actually assessed would be limited by the testing time available, and the need to adequately support inferences about the extent of each student's mastery of each assessed aim (what they termed an "all concepts eligible, some concepts tested" approach).

In terms of the tests themselves, they proposed that tests should be administered so that teachers had time to use the outcomes to inform their instruction, that the tests should include at least two classroom assessments for each curricular aim, and that the tests should be principally focused on mastery of concepts, rather than skills. The group recognized that, for reasons of economy, it would be necessary for the majority of items to be in selected-response format, but recommended that a modest proportion of the items should require constructed responses. To enable teachers to derive optimal instructional insights from the tests, the group also recommended that teachers should be provided with relevant professional development. Finally, in addition to the tests of science achievement, students should complete anonymously completed self-report inventories that teachers can use to gauge their students' science-related attitudes and interests.

As appendices to their report, the group offered a number of practical suggestions for taking this work forward. One appendix provided a number of illustrative items, each closely related to a particular cell in the "skill-by-concept" matrix discussed above. For example, for the cell that required students to develop descriptions, explanations, predictions, and models in the context of characteristic chemical properties and changes, the following item was proposed:

> "Iron rusting can be explained as a chemical change. What evidence can be used to support this explanation?" (p. 152)

A second appendix included an illustrative "Request for proposals" (RFP) that a state might release to potential contractors interested in providing instructionally supportive accountability tests in science. As well as the obvious test specifications, the RFP requires contractors to identify instructional suggestions for the teaching of areas that the level of performance on the test indicates are causing difficulties for students, how teachers are to be involved in the construction of the tests, and how students with special needs are to be assessed.

Until such proposals are fully implemented in a state-wide testing context, whether these ideas are workable or not remains to be seen. There is also the question of whether such tests would really be more sensitive to the effects of instruction. At first sight, it would appear that they should be, in that they would relate very closely to a small number of curricular aims that would be, for that year, the clear focus of instruction, but again, this remains to be seen. However, the suggestions made by Popham *et al*. seem to be the most carefully thought through proposals for the creation of a system of high-stakes accountability tests that, as well as providing information about the levels of achievement in schools, could, at the same time, help teachers improve instruction. As such, they deserve serious consideration by all those involved in high-stakes accountability testing.

Conclusion

The conclusion of this article is somewhat paradoxical. Because differences between schools account for only a small proportion of the variance in student scores (in most

countries, less than ten percent), standardized tests are rather inappropriate tools with which to hold districts, schools and teachers accountable. And yet, there is evidence that establishing an accountability regime that uses externally-set tests, where the results of these tests have significant consequences for students, teachers, schools and districts, can be a cost-effective way to increase student achievement, although the introduction of such regimes has the potential for a range of unintended outcomes, many of which will have a negative impact.

The research reviewed in this article suggests that there is a case for the use of high-stakes accountability tests, but that considerable work needs to done to minimize the costs and maximize the benefits of such regimes. This is a challenging agenda because, as Popham *et al*. (2005) acknowledge, it will require integrating a range of disciplinary perspectives, including economics, psychometrics, psychology, subject expertise, and knowledge of teacher professional development. Ultimately if we are to have high-stakes tests, the search must be for "tests worth teaching to" (Resnick, 1987): accountability tests that are so closely aligned with desired outcomes that the only way to improve scores is to improve the desired outcomes.

References

Amrein, A. L., & Berliner, D. C. (2002b). *The impact of high-stakes tests on student academic performance: An analysis of NAEP results in states with high-stakes tests and ACT, SAT, and AP test results in states with high school graduation exams*. Retrieved December 2002, from Educational Policy Studies Laboratory, Education

Policy Research Unit: http://epicpolicy.org/files/EPSL-0211-126-EPRU.pdf

Amrein, A.L. & Berliner, D.C. (2002a, March 28). High-stakes testing, uncertainty, and student learning *Education Policy Analysis Archives*, *10*(18). Retrieved June 1, 2009 from http://epaa.asu.edu/epaa/v10n18.

Bardach, E., & Lesser, C. (1996). Accountability in human services collaboratives—for what? And to whom? *Journal of Public Administration Research and Theory,* **6**(2), 197-224.

Bishop, J. H. (2001a). A steeper, better road to graduation. *Education Next,* **1**(4), 56-61.

Bishop, J. H. (2001b). *Why do students learn more when achievement is examined externally?* Retrieved June 11, 2009, from http://media.hoover.org/documents/ednext20014unabridged_bishop.pdf.

Black, P., & Wiliam, D. (2005). Lessons from around the world: how policies, politics and cultures constrain and afford assessment practices. *Curriculum Journal,* **16**(2), 249-261.

Braun, H. (2004, January 5). Reconsidering the impact of high-stakes testing, *Education Policy Analysis Archives, 12*(1). Retrieved June 1, 2009 from http://epaa.asu.edu/epaa/v12n1/.

Braun, H. I. (2005). *Using student progress to evaluate teachers: a primer on value-added models*. Princeton, NJ: Educational Testing Service.

Brown, M. L. (Ed.). (1992). *Graded Assessment in Mathematics: teacher's guide*. Walton-on-Thames, UK: Nelson.

Brown, M. L., Blondel, E., Simon, S. A., & Black, P. J. (1995). Progression in measuring. *Research Papers in Education,* **10**(2), 143-170.

Campbell, D. T. (1976). *Assessing the impact of planned social change* (Vol. 8). Hanover, NH: The Public Affairs Center, Dartmouth College.

Cannell, J. J. (1988). Nationally normed elementary achievement testing in America's public schools: How all fifty states are above the national average. *Educational Measurement: Issues and Practice,* **7**(2), 5-9.

Carneiro, P., Crawford, C., & Goodman, A. (2007). *Impact of early cognitive and non-cognitive skills on later outcomes*. London, UK: London School of Economics Centre for the Economics of Education.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis,* **24**, 305-331.

Chambers, J. G., Parrish, T. B., Levin, J. D., Smith, J. R., Guthrie, J. W., Seder, R. C., & Taylor, L. (2004). *The New York adequacy study: determining the cost of providing all children in New York an adequate education*. Washington, DC: American Institutes for Research/Management Analysis and Planning.

Cizek, G. J. (2005). High-stakes testing: contexts, characteristics, critiques, and consequences. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 23-54). Mahwah, NJ: Lawrence Erlbaum Associates.

Clotfelter, C. T., Ladd, H. F., Vigdor, J. L., & Diaz, R. A. (2003). *Do school accountability systems make it more difficult for low performing schools to attract and retain high quality teachers?* Paper presented at the Annual Meeting of the American Economic Association held at Washington, DC.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Committee of Inquiry into the Teaching of Mathematics in Schools. (1982). *Report: mathematics counts*. London, UK: Her Majesty's Stationery Office.

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence,* **35**(1), 13-21.

Denvir, B., & Brown, M. L. (1986a). Understanding of number concepts in low-attaining 7-9 year olds: part 1. Development of descriptive framework and diagnostic instrument. *Educational Studies in Mathematics,* **17**(1), 15-36.

Denvir, B., & Brown, M. L. (1986b). Understanding of number concepts in low-attaining 7-9 year olds: part II. The teaching studies. *Educational Studies in Mathematics,* **17**(2), 143-164.

Educational Testing Service Cooperative Test Division. (1957). *Cooperative Sequential Tests of Educational Progress: technical report*. Princeton, NJ: Educational Testing Service.

Feinstein, L., Budge, D., Vorhaus, J., & Duckworth, K. (2008). *The social and personal benefits of learning: a summary of key research findings*. London, UK: Institute of Education, University of London.

Foxman, D. D., Cresswell, M. J., Ward, M., Badger, M. E., Tuson, J. A., & Bloomfield, B. A. (1980). *Mathematical development: primary survey report no 1*. London, UK: Her Majesty's Stationery Office.

Foxman, D. D., Martini, R. M., Tuson, J. A., & Cresswell, M. J. (1980). *Mathematical development: secondary survey report no 1*. London, UK: Her Majesty's Stationery Office.

Goldin, C. (2002). American leadership in the human capital century: have the virtues of the past become the vices of the present? In Y. K. Kodrzycki (Ed.), *Education in the 21st Century: meeting the challenges of a changing world* (pp. 25-35). Boston, MA: Federal Reserve Bank of Boston.

Haertel, E. H., & Herman, J. L. (2005). A historical perspective on validity arguments for accountability testing. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement: 104th Yearbook of the National Society for the Study of Education (part 2)* (pp. 1-34). Malden, MA: Blackwell.

Hanushek, E. A. (Ed.). (2006). *Courting failure: how school finance lawsuits exploit judges' good intentions and harm our children*. Stanford, CA: Hoover Institution.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management,* **24**(2), 297-327.

Harlen, W., & Deakin-Crick, R. (2002). A systematic review of the impact of summative assessment and tests on students' motivation for learning. In EPPI-Centre (Ed.), *Research Evidence in Education Library* (1.1 ed., pp. 153). London, UK: Institute of Education Social Science Research Unit.

Hart, K. M. (Ed.). (1981). *Children's understanding of mathematics: 11-16*. London, UK: John Murray.

Heritage, M. (2008). *Learning progressions: supporting instruction and formative assessment*. Washington, DC: Council of Chief State School Officers.

Herman, J. L., Abedi, J., & Golan, S. (1994). Assessing the effects of standardized testing on teaching in schools. *Educational and Psychological Measurement,* **54**(2), 471-482.

Hieronymous, A. N., & Lindquist, E. F. (1974). *Manual for administrators, supervisors and counselors – levels edition (forms 5 &6): Iowa tests of basic skills*. Boston, MA: Houghton Mifflin.

Hodgen, J., Küchemann, D. A., Brown, M. L., & Coe, R. (2009). *Secondary students' understanding of mathematics 30 years on*. Paper presented at the Annual meeting of the British Educational Research Association. Manchester, UK.

Hurt, John (1971). *Education in Evolution: Church, State, Society and Popular Education*

1800-1870. London: Rupert Hart-Davis.

Jesson, D., & Crossley, D. (2007). *Educational outcomes and value added by specialist schools: 2006 analysis*. London, UK: Specialist Schools and Academies Trust.

Kent, P., & Blows, D. (2009). *Understanding the CVA model* (Vol. 58). Leicester, UK: Association of School and College Leaders.

Kellner, P. (1997, 19 September). Hit-and-miss affair. *Times Educational Supplement,* 23.

Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April). *The effects of high-stakes testing: preliminary evidence about generalization across tests.* Paper presented at the Annual meetings of the American Educational Research Association and the National Council on Measurement in Education held at Chicago, IL.

Koretz, D. M., Stecher, B. M., Klein, S. P., McCaffrey, D., & Deibert, E. (1994). *Can portfolios assess student performance and influence instruction? The 1991-92 Vermont experience* (Vol. RP-259). Santa Monica, CA: RAND Corporation.

Lee, J. (2008). Is test-driven external accountability effective? Synthesizing the evidence from cross-state causal-comparative and correlational studies. *Review of Research in Education,* **78**(3), 608-644.

Levin, H. M., Belfield, C., Muennig, P., & Rouse, C. (2007). *The costs and benefits of an excellent education for all of America's children*. New York, NY: Teachers College.

Mackintosh, N. J. (2000). *IQ and human intelligence*. Oxford, UK: Oxford University Press.

Madaus, G. F. (Ed.). (1983). *The courts, validity and minimum competency testing*. Boston, MA: Kluwer Academic Publishers.

McGaw, B. (2008). The role of the OECD in international comparative studies of achievement. *Assessment in Education: Principles Policy and Practice,* **15**(3), 223–243.

Mencken, H. L. (1917, 16 November). The Divine Afflatus*. New York Evening Mail*. Reprinted in H. L. Mencken (Ed.), *A Mencken Chrestomathy* (pp. 443). New York, NY: Alfred A. Knopf.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3 ed., pp. 13-103). Washington, DC: American Council on Education/Macmillan.

Mitch, D. (1999, November 12). *Social accountability and educational outcomes: interpreting the episode of payment by results in Victorian England*. Paper presented at a meeting of the Social Science History Association held at Fort Worth, TX.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment frameworks*. Chestnut Hill, MA: Boston College Lynch School of Education.

National Assessment of Educational Progress. (2006). *The nation's report card: Mathematics 2005* (Vol. NCES 2006-453). Washington, DC: Institute of Education Sciences.

Newmann, F. M., Bryk, A. S., & Nagaoka, J. K. (2001). *Authentic intellectual work and standardized tests: conflict or coexistence?* Chicago, IL: Consortium on Chicago School Research.

Norcini, J. J. (2009). Personal communication (November 16).

Organisation for Economic Cooperation and Development. (2000). *Measuring student knowledge and skills: the PISA 2000 assessment of reading, mathematical and scientific literacy*. Paris, France: Organisation for Economic Cooperation and Development.

Organisation for Economic Cooperation and Development. (2004). *Learning for tomorrow's world: first results from PISA 2003*. Paris, France: Organisation for Economic Co-operation and Development.

Organisation for Economic Cooperation and Development. (2007). *Science competencies for tomorrow's world, volume 1: analysis*. Paris, France: Organisation for Economic Co-operation and Development.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming and equating. In R. L. Linn (Ed.), *Educational measurement* (3 ed., pp. 147-200). Washington, DC: American Council on Education/Macmillan.

Phelps, R. P. (2005). The rich, robust research literature on testing's achievement benefits. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 55-90). Mahwah, NJ: Lawrence Erlbaum Associates.

Pipho, C. (2002). *Seven lessons learned from minimum competency testing*. Denver, CO:

Education Commission of the States.

Popham, W. J., Keller, T., Moulding, B., Pellegrino, J. W., & Sandifer, P. (2005). Instructionally supportive accountability tests in science: a viable assessment option? *Measurement: Interdisciplinary Research and Perspectives,* **3**(3), 121-179.

Rapple, B. A. (1994). Payment by results: an example of assessment in elementary education from nineteenth century Britain. *Education Policy Analysis Archives,* **2**(1). Retrieved June 1, 2009 from http://epaa.asu.edu/epaa/v2n1/.

Ray, A. (2006). *School value added measures in England: a paper for the OECD project on the development of value-added models in education systems*. London, UK: Department for Education and Skills.

Resnick, L. B. (1987). *Education and learning to think*. Washington, DC: National Academy Press.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response Items: a random effects synthesis of correlations. *Journal of Educational Measurement,* **40**, 163-184.

Rodriguez, M. C. (2004). The role of classroom assessment in student performance on TIMSS. *Applied Measurement in Education,* **17**(1), 1-24.

Rosenshine, B. V. (2003). High-stakes testing: another analysis. *Education Policy Analysis Archives,* **11**(24).

Royal Commission. (1861). *Report (Newcastle Report)* (Vol. HC 1861 xxi). London: Her Majesty's Stationery Office.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: searching for instructional sensitivity. *Journal of Research in Science Teaching,* **39**(5), 369-393.

Spearman, C. (1904). "General Intelligence" objectively determined and measured. *American Journal of Psychology,* **15**, 201-293.

Sylvester, D. W. (1974). *Robert Lowe and Education*. Cambridge, UK: Cambridge University Press.

Wescott, J. P. (1972, February 12-16). *Accountability: for whom, to whom, for what?* Paper presented at the 104th annual meeting of the American Association of School Administrators held at Atlantic City, NJ. Retrieved September 1, 2009, from http://www.eric.ed.gov:80/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/3b/0f/cc.pdf

White, E. E. (1886). *The elements of pedagogy*. New York, NY: American Book Company.

Wibowo, A., Hendrawan, I., & Deville, C. (2009, 16 April). *Design of the vertical scale: test development, data collection design, linking design*. Paper presented at the annual meeting of the American Educational Research Association held at San Diego, CA.

Wiliam, D. (1992). Value-added attacks? Technical issues in publishing national

curriculum assessments. *British Educational Research Journal,* **18**(4), 329-341.

Wiliam, D. (2007). Once you know what they've learned, what do you do next? Designing curriculum and assessment for growth. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school* (pp. 241-270). Maple Grove, MN: JAM Press.

Wiliam, D. (2010). What counts as evidence of educational achievement? The role of constructs in the pursuit of equity in assessment. In A. Luke, G. Kelly & J. Green (Eds.), *What counts as evidence in educational settings? Rethinking equity, diversity and reform in the 21st century* (Vol. 34). Washington, DC: American Educational Research Association.

Willms, J. D. (1992). *Monitoring school performance: a guide for educators*. London, UK: Falmer.

Wood, R. (1991). *Assessment and testing: a survey of research*. Cambridge, UK: Cambridge University Press.

---

[1] Indeed, at the heart of the legislation there is a perverse incentive for students to perform badly on the tests, because if the school fails to make adequate yearly progress towards the goal of proficiency for all students, then parents get additional rights to supplementary education and choice of schools.

[2] The metric used for reporting CVA scores for secondary schools has a mean of 1000 with a score of 1048 representing a school in which on average students scored one grade higher in each of the eight included subjects (e.g., so that a student gains 8 grade

Bs in such a school rather than eight grade Cs in an average school). The standard deviation of the CVA scores in 2007 was 17, equivalent to 0.35 grade points.

Figure

Click here to download Figure: Figure 1.doc

Figure 1: Increase in facility of an arithmetic item with age

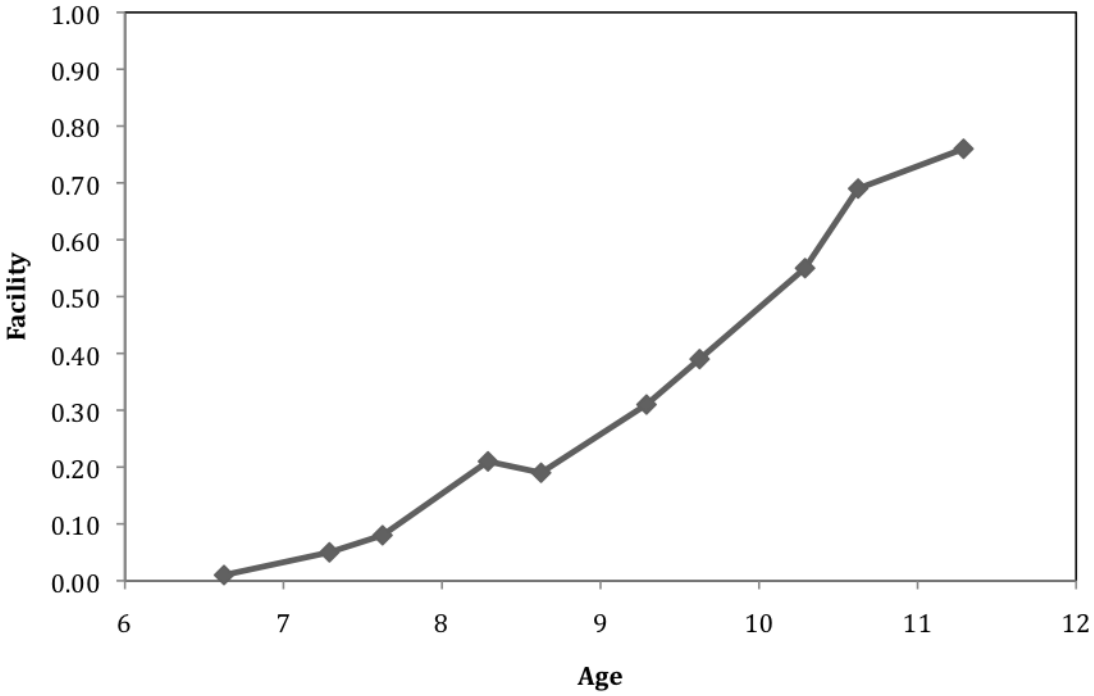

Figure 2
Click here to download Figure: Figure 2.doc

Figure 2: Achievement in Decimals by age found in CSMS (Hart, 1981)

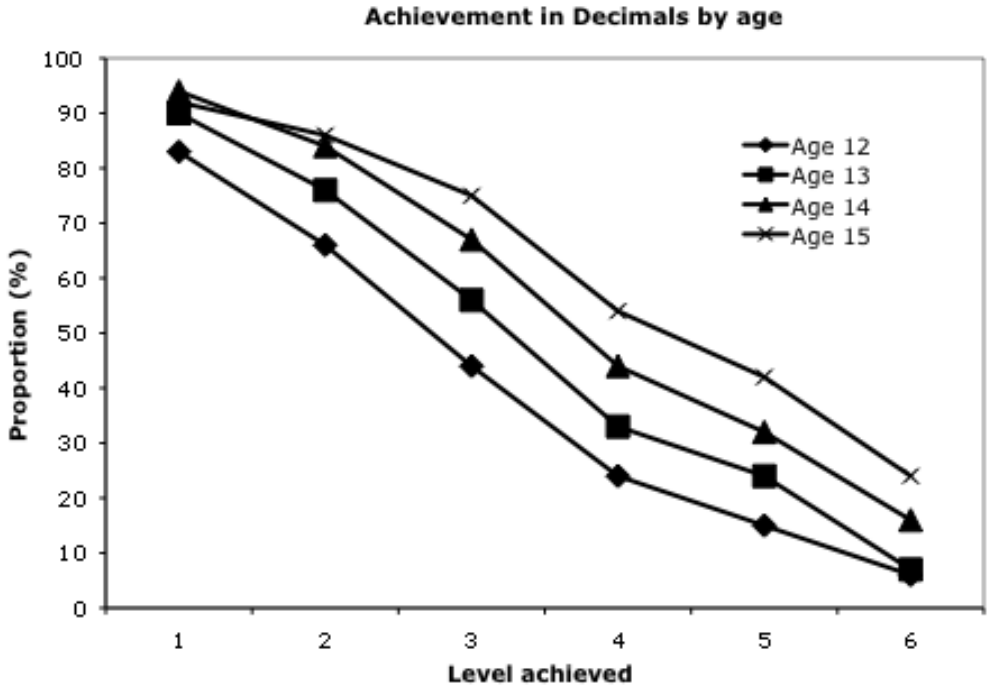

Figure 3
Click here to download Figure: Figure 3.doc

Figure 3: Annual growth in school attainment in the ETS STEP tests (1957)

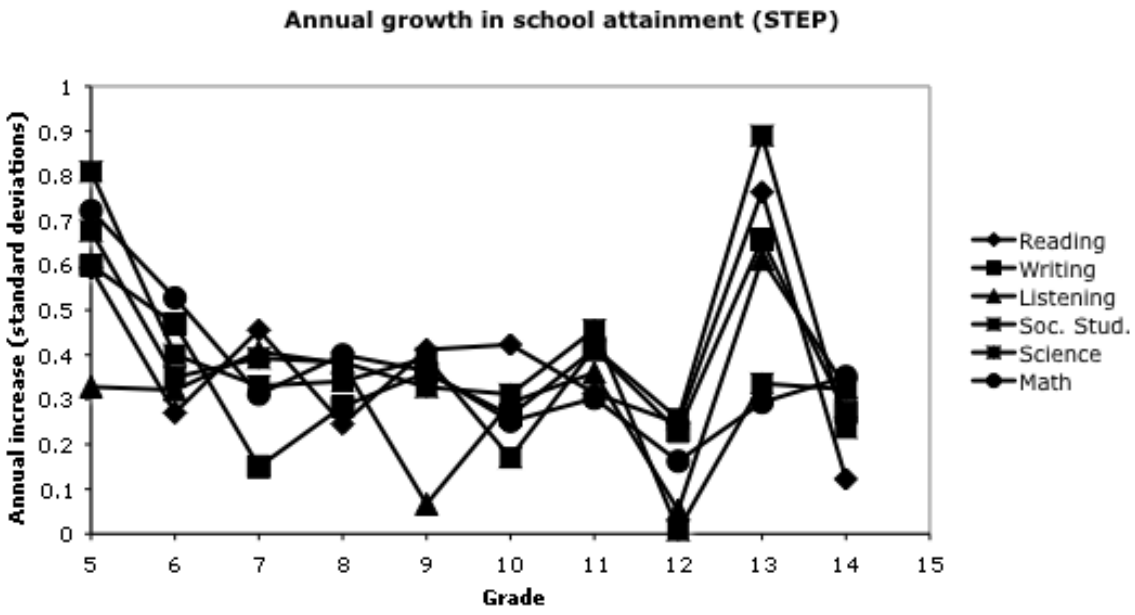

Figure 4
Click here to download Figure: Figure 4.doc

Figure 4: Annual growth in school attainment in Connecticut