Sensitivity to instruction: the missing ingredient in large-scale assessment systems?

Paper presented at the annual meeting of the International Association for Educational Assessment: Baku, Azerbaijan, September 2007

Dylan Wiliam
Institute of Education, University of London

Abstract

Educational policymakers all over the world rely on the results of large-scale accountability tests to inform policy. In doing so, they assume that the scores obtained by students indicate the quality of instruction the students have received—i.e., that the tests are sensitive to instruction. Using data from a variety of tests and examinations, this paper will establish that typical standardized tests are in fact not at all sensitive to instruction, for two reasons. The first is that the progress made by individual students is actually far less than the variability within an age cohort. The second reason is that the traditional processes of test construction decrease the sensitivity of a test to instruction, by systematically eliminating items that *are* sensitive to instruction. The paper concludes with two policy-relevant measures to address this. The first is that we should change the way we calculate reliability coefficients to prevent the systematic exclusion of items that are sensitive to instruction, and the second is a public information campaign to raise awareness of the issue of sensitivity to instruction, so that users of accountability test results understand the limitations of these tests as measures of the quality of education provided.

Increases of item facility with age

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 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 for a considerable time. One item (code 1106) was presented orally to students in first through fifth grade and asked students to complete the following calculation:

“Eight hundred and sixty add five hundred and seventy”[1]

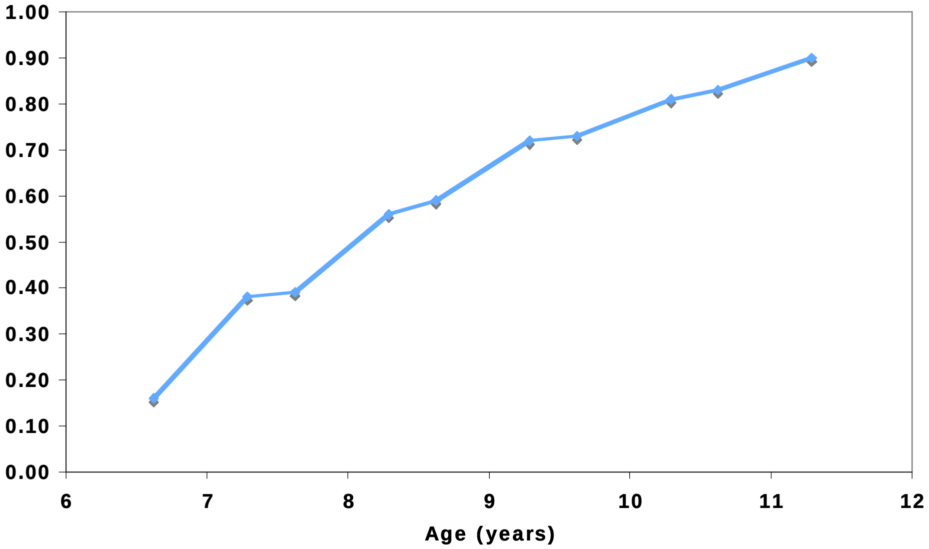
Although the item was presented orally, students were provided with scratch paper for calculation, and were asked to write their answer in a space given to them in test booklet. The facility of the item for students of different ages is as shown in figure 1.

Perhaps the most remarkable feature of figure 1 is how slowly the facility of this item increases with age. Twenty percent of students can answer the item correctly well

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

before the age of 7, but three years later, twenty percent still cannot. The testing of the students was undertaken as part of a research program, and thus was low-stakes for both teachers and students, and was not keyed to any curricular coverage of this skill. Since this was a nationally representative sample, it is reasonable to assume that this is a reasonably accurate indication of the "response to treatment" under typical instruction. One explanation for the relative slowness of the increase in facility could be that this was not a skill that teachers taught. However, in a teacher questionnaire, all teachers were asked to indicate whether the skills being assessed were skills they sought to develop in students, and almost all teachers (well over 90%) indicated that this 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 it is presence in school, rather than general maturation, that is the main cause of the increase in facility.

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



Of course this is just a single isolated item and it might be argued that other items would show greater increases in facility, 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. This is rather astonishing. 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 the teachers 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. And yet, in a class of 24 students, an annual increase of twenty-five percentage points means that only six students would acquire the skill that year. The other 18 students in the class would 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

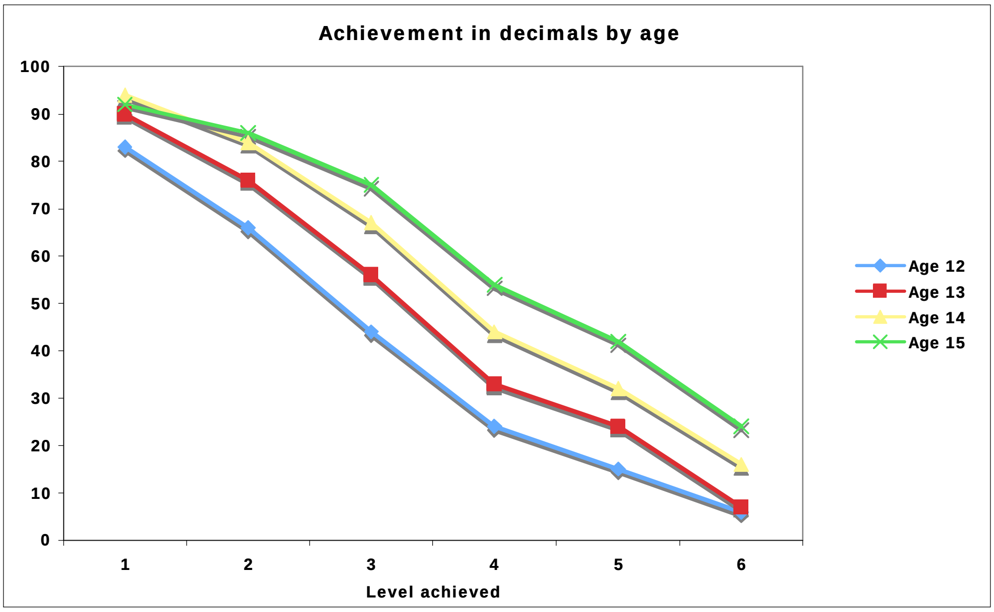
USA) in 1980 had shown that high-achieving 7-year-old students out-performed some 14-year-olds on basic arithmetic. One item in particular:

$$6099 + 1 = ?$$

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). In fact, it was this item that led to the idea that there was 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).

General competences in mathematics also showed the same, or even greater, variability that had been found by the APU. The Concepts in Secondary Mathematics and Science (CSMS) project had identified a series of 6 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, the proportion of students achieving a particular level increased by only 5-10% per year (see Figure 2).

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



Brown, Blondel, Simon and Black (1995) found that in England performance on some of the more conceptual issues involved in measuring length and weight (mass) were also slow to change. 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).

Deleted: *et al.*

Deleted: the performance of some students in the conceptual issues

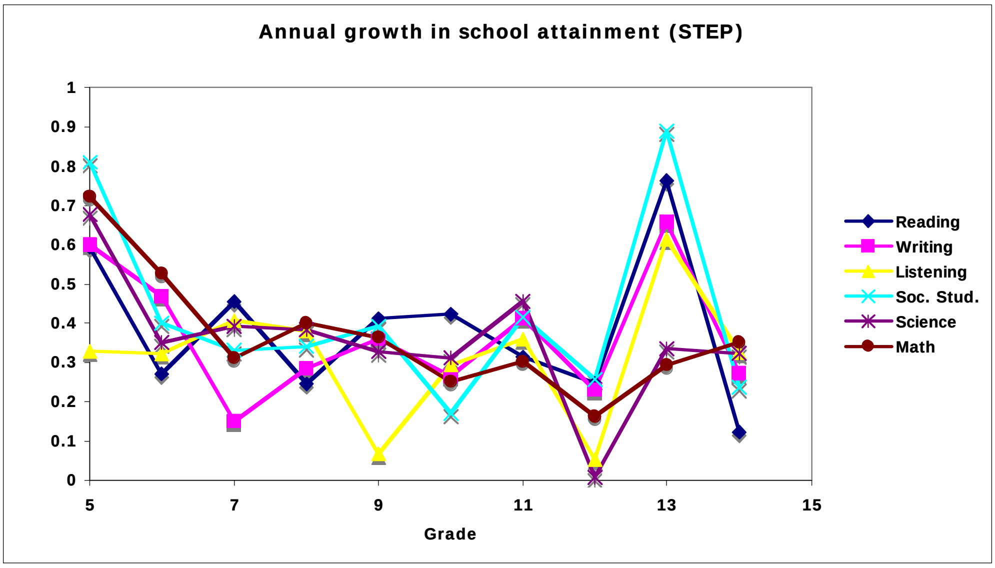
Deleted: in

Deleted: was

Deleted: ’

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, suggesting that a student at the 95th percentile is as much as ten years ahead of a student at the 5th percentile.

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



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. However, some of the higher attaining students will have achieved a higher level. The data from the ITBS scaling studies indicate that about 30% of students will, by half way through grade 3, have achieved an attainment 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. Collecting similar data points for third graders and joining in them up gives us 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. While Petersen *et al.* advocate caution in making interpretations about the equivalence of performance of students of different age, it is clearly the case that in some sense, even in language usage, some third-graders are performing like 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 many current measurements of the annual growth in achievement still typically range 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.

Why should one standard deviation of achievement be equivalent to four year's growth in the NAEP mathematics and reading tests, less than three years growth on the TIMSS mathematics tests, and around two years' growth on the fourth-grade ITBS? The answer is, at least in part, construct definition (Braun, Jackson & Wiley, 2001). The same subject can be defined in a number of ways, and these different choices give rise to different properties, not least in the spread of achievement.
For example, when ability in science is defined in terms of scientific reasoning, perhaps best exemplified by the science reasoning tasks developed by Shayer and Adey (1981), achievement is less closely tied to age and curriculum exposure, and more closely related to measures of general reasoning (Shayer, Kücheman & Wylam, 1976; Shayer & Wylam). In other words, the science reasoning tasks are not strongly related to quality of instruction received, or maturation. In contrast, when science is defined in terms of knowledge of facts that are taught in school, then opportunity to learn will be the most important factor—those students who have been taught the facts will know them, or at least have the opportunity to know them, while those who have not been taught will, in all probability, not. A test that assesses these skills is likely to be highly influenced by the amount and quality of instruction. A third case might arise in the discussion of ethical and moral dimensions of science, where maturity, rather than general intelligence or curriculum exposure, might be the most important factor. Here it might well be that a student's performance depends relatively little on the amount or quality of instruction, or on general intelligence, but is highly sensitive to maturation.

The fundamental question here is what changes a student's score? From this perspective, sensitivity to instruction—the focus of this symposium—is just one type of sensitivity. A test is sensitive to instruction when instruction changes a student's score. If instruction does not change a student's score on a test very much, then that test is insensitive to instruction, but it may be sensitive to maturation. We might find that on an item or a test, students' scores change little after instruction, but a year later, even in the absence of instruction, the scores may be significantly higher. This item or test would be insensitive to instruction, but sensitive to maturation. A third case would be when a student's score on a test is changed little either by maturation or high quality instruction, but is strongly linked to measures of general ability, such as the science reasoning tasks discussed above.

Before we can examine the policy implications of these findings, however, there are two additional complications to consider; the spread of achievement within the cohort is not necessarily constant over time, and it may be affected by the way in which tests of achievement are constructed. These two issues are discussed in turn below.

Does variability increase with age?

Over a century and a half ago Quetelet (1835) noted that as the heights of Belgian boys and girls increased from birth to adulthood, so did the range of heights within a cohort of students of the same age, and the correlation of means with standard deviations appears to be almost universal, at least for physical measurements. For cognitive measurements, the picture is more complex. Time-indexed measures (e.g. by grade or age) do show much the same picture (see for example, Williamson, Applebaum, & Epanchin, 1991; Wiliam, 1992). On the other hand, it has been found that some order-preserving measures (such as those that use item-response modeling) do not (Yen, 1986). Part of the reason for this is that item-response models used in many of these studies have fitted models to populations grade-by-grade, which has the effect of constraining the spread within each grade. In general, while item-response models do show lower rates of increase of spread, they do also generally show some increase (although there are some ceiling effects). However, the most important finding of this research is that there is no ‘natural’ way of measuring the growth of achievement over time, and that different models have different properties.

Deleted: In 1835,
Deleted: et al.
Deleted: most
Deleted: ‘
Deleted: ’

The issue of growth models is important, because it has profound implications for the distribution of attainment within a cohort, and the rate at which achievement increases over time. To see why, consider figure 4. It shows a hypothetical distribution of achievement across cohorts aged from 4 to 16 years of age with achievement being measured in “attainment age”. A 14-year-old student would have an attainment age of 12 if his or her achievement were equivalent to that of the average 12-year-old. Of course, whether there are such 14-year-olds is an empirical question, and depends on a range of factors, including the extent to which the domain can be regarded as unidimensional. The model shown in figure 4 assumes that the standard deviation of achievement within a cohort is one-tenth of the chronological age.

Deleted: choice
Deleted: “
Deleted: ”

The importance of the relationship between standard deviation and age becomes apparent if we look at figure 5, which shows the distribution of achievement if the standard deviation of achievement is one-fifth of the chronological age. The peak of the distribution remains in the same place of course, but the amount of overlap between cohorts is much greater in figure 5 than in figure 4, and greater again in figure 6, which shows what happens when the spread is one-fourth the chronological age.

It seems likely that the view of most stakeholders in education would be that even figure 4 *understates* the differences between age cohorts. Most parents, politicians, and even many educators, seem to assume that the achievement of almost all students in eighth grade would be higher than that of almost all students in seventh grade. In fact, empirical evidence suggests that figures 5 and 6 are the most realistic.

Wiliam (1992) analyzed tables of norms published for a range of published tests in order to investigate the rate of increase of variability with age. These norms have been derived through sophisticated processes involving a number of assumptions that can make interpreting them difficult. On the other hand, such tables do represent claims about the relative performance of students of different ages. For example, if the average raw score for an eight-year-old in a particular test were 38, then using a score scale with mean of 100, and standard deviation 15, this raw score of 38 would be reported as a standardized score of 100 for age 8. If the mean raw score for ten year-olds were 46,

Figure 4: distribution of achievement with SD at one-tenth of chronological age

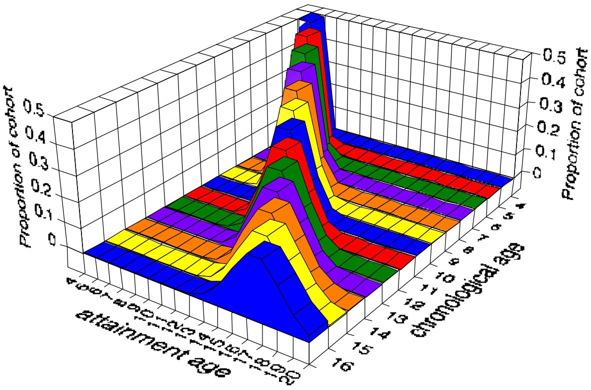


Figure 5: distribution of achievement with SD at one-fifth of chronological age

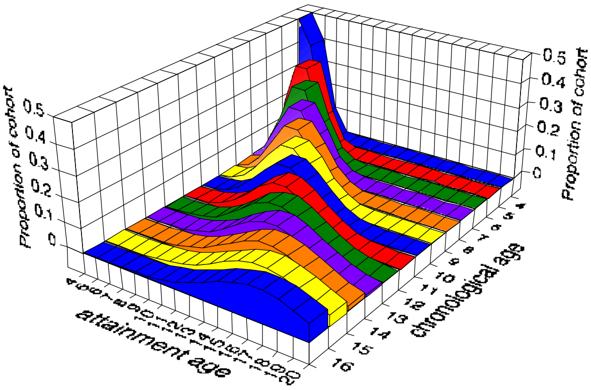
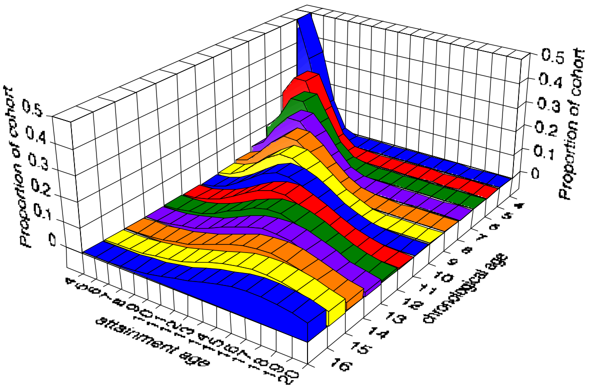


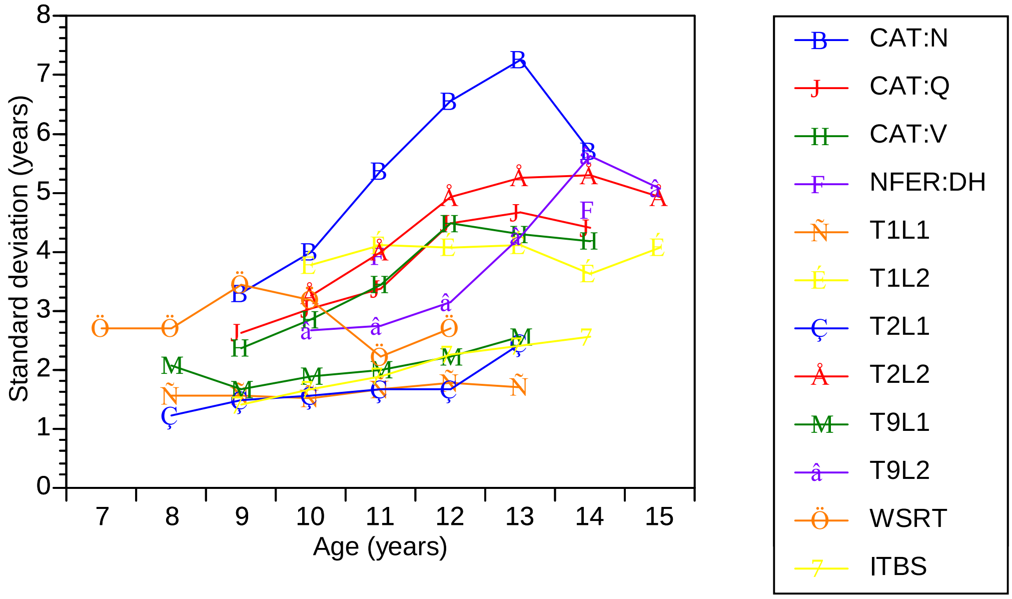
Figure 6: distribution of achievement with SD at one-fourth of chronological age



and this same raw score were associated with a standardized score of 115 for 8-year-olds, then this represents a claim by the test constructors that 8-year-old students who are one standard deviation above the mean, are, in a very real sense, two years ahead of average for this test.

The result of aggregating data from such comparison across the sets of norms made available by the publishers of eleven tests of mathematics, reading and writing for upper elementary and secondary school students is shown in figure 7. A similar analysis for the data from ITBS that was discussed above is included in figure 7 for reference. The conclusion that variability does increase with age would appear to be strongly supported by these data, with some tests (like ITBS) indicating that the standard deviation is around one-fifth the chronological age, and others indicating rates up to one-third, or even approaching half the chronological age.

Figure 7: Rate of increase of spread of achievement in 11 standardized tests (see text)



Key:

CAT:N	Cognitive abilities test non-verbal battery (Thorndike et al, 1986)
CAT:Q	Cognitive abilities test quantitative battery (Thorndike et al, 1986)
CAT:V	Cognitive abilities test verbal battery (Thorndike et al, 1986)
CSMS(M)	Concepts in Secondary mathematics and Science (Hart, 1980)
NFER:DH	NFER non-verbal test DH (Calvert, 1958)
T1L1	Profile of mathematical skills test 1 level 1 (France, 1979)
T1L2	Profile of mathematical skills test 1 level 2 (France, 1979)
T2L1	Profile of mathematical skills test 2 level 1 (France, 1979)
T2L2	Profile of mathematical skills test 2 level 2 (France, 1979)
T9L1	Profile of mathematical skills test 9 level 1 (France, 1979)
T9L2	Profile of mathematical skills test 9 level 2 (France, 1979)
TGAT	Task Group report (National Curriculum Task Group on Assessment and Testing, 1987)
WSRT	Wide-span reading test (Brimer, 1984)

These differences are not, primarily, related to the subject being assessed. In the profile of Mathematical Skills tests (France, 1979) the level 1 tests show modest increases in variability with increasing age, while the level 2 tests show much greater increase of spread. It does appear, however, that skills-based tests have a slower rate of increase of spread of achievement than reasoning-based tests, with the greatest spread, and the sharpest increase, perhaps not surprisingly, being shown by the non-verbal reasoning battery of the Cognitive Skills Test (Thorndike et al., 1986). Construct definition therefore appears to impact not only the spread of achievement, but also its rate of increase with age, although whether these two are independent or not requires further investigation, which is beyond the scope of this paper.

Test construction methods

It hardly needs saying that an adequate degree of reliability is essential for any assessment, but it is less widely understood that efforts to increase reliability can change the construct that the test is measuring. One can think of the classical reliability coefficient as a kind of signal to noise ratio (or more accurately as a signal to signal-plus-noise ratio). It is therefore possible to improve the reliability by decreasing the noise *or by increasing the signal*. This is why test developers seek items that discriminate between candidates, for they increase the signal, thus improving the reliability. In consequence, items that all students answer correctly, or ones that all students answer incorrectly, are generally omitted, since they do not discriminate between students, and thus do not contribute to reliability. This alters the construct being measured by the test, because when we develop a test for students in, say, the eighth-grade, it is customary to trial the test only with eighth-grade students. The result is that items that discriminate between eighth-grade students are retained, and those that do not are not. To see why is this so important consider what would happen to an item that no seventh-grade student can answer correctly, but can be answered correctly by all eighth-grade students. This item is almost certainly assessing something that is changed by instruction. And yet with traditional test development processes, the item would be retained neither in a test for seventh graders, nor one for eighth graders. It would not be retained in a test for seventh graders because it is too hard, while it would not be retained in a test for eighth graders, because it is too easy. In neither grade does the item discriminate between students in the same grade, even though it does discriminate well between seventh graders and eighth graders. Such items are therefore routinely omitted from tests. The reliability of the test is increased, but the extent to which the test measures the effects of instruction is reduced.

Formatted: Font: Italic

Summing up the argument so far, I have shown that student achievement is relatively insensitive to instruction, in the sense that under conditions of average instruction, the progress of students is slow compared to the variability of achievement within the age cohort. I have shown that in many, if not most, of the measures that are in use across multiple grades, the spread of achievement increases with age, and that the standard techniques of test construction are likely to reduce the extent to which tests measure those things that are changed by instruction.

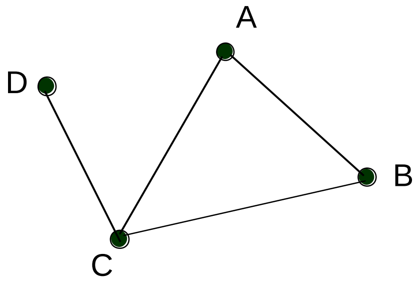
Implications for policy

Why does any of this matter? It matters because if we are to hold schools to account, we should hold them to account for the things that they can change. The analysis above shows that, in a very real sense, learning in school is resistant to instruction. In terms of the three “dimensions of progression” described above (instruction, maturation, general reasoning), only the quality of instruction (although conceived in its broadest sense) is under the school's control, and therefore, schools should be held accountable only for the quality of instruction. The question is how to do this.

One policy response might be to change the construct assessed in school subjects to focus more on the aspects of each subject that are impacted by instruction and less on the uncontrollable dimensions, such as general ability or maturation. For example, we could fill our mathematics curriculum with topics like matrices and networks. The basic idea here is that a network of nodes and arcs can be represented by a matrix in which ones represent an arc between two nodes and zeroes represent the absence of such arcs, as shown below in figure 7 below.

Figure 7: hypothetical item highly sensitive to instruction

$$\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$



Items on this topic are strongly linked to curriculum exposure. The skills are not difficult to teach, in that most students in a class can learn the skills relatively easily once they are shown the basic idea, but without being shown what to do, even students with strong general reasoning skills are generally unable to work out what is required. Tests full of such items would display a high degree of sensitivity to instruction, but at a terrible price. We would have made our tests sensitive to instruction by discarding all elements of a subject where critical thinking, reasoning, and the ability to make connections confer an advantage. Indeed, there is evidence that the United States has already gone further down this road than other developed countries. The correlation of IQ with school achievement is approximately 0.7 in the United Kingdom, but around 0.4 in the United States (Mackintosh, 2000), possibly attributable to US curricula that are "a mile wide and an inch deep" (Schmidt, McKnight, Raizen, 1997 p. 62).

If we cannot change the curriculum, what else can we do? One straightforward measure would be to ensure that the processes of test construction do not reduce a test's sensitivity to instruction any further. It seems obvious that to develop a test intended for eighth graders, we should calculate reliability on the basis of the scores of eighth graders. Indeed, it would be seen as highly unethical to increase the reported reliability of a test artificially by increasing the true-score variance by the addition of the scores of students for whom the test was not appropriate (e.g., fifth graders and eleventh graders). However, if a test is to be used to assess the quality of eighth-grade instruction, for example, it does not seem unreasonable to calculate reliability on the basis of those who have received instruction and those who have not. Ideally, these would be eighth graders who have not received instruction on the constructs addressed by the test, but this may be difficult to arrange. For that reason, it seems reasonable that reliability measures for accountability tests for eighth graders should be derived from samples of seventh and eighth graders.

However, since as argued above, it is really achievement, rather than just tests, that are most insensitive to instruction, such measures are likely to have only a limited impact. What I propose instead, is a campaign of public education to communicate to stakeholders the limited impact that instruction can have on achievement.

At first sight, this claim seems counter-intuitive. There is an emerging literature that shows that teacher effects are generally much greater than school effects (Wright, Horn & Sanders, 1997). In other words, it matters far less which school you go to than which teachers you get in school. For example, according to Hanushek (2002), a teacher at the 95th percentile of teacher quality generates student achievement at twice the rate of the average teacher, and a teacher at the 5th percentile generates student achievement at half the rate of the average teacher. This seems an impressive difference, but if we assume that the average increase in achievement per year is 0.3 standard deviations, then high-quality instruction will add only 0.6 standard deviations, and very low quality instruction will add 0.15 standard deviations. To put this into perspective, this difference means that a class of one of the very best teachers will have three more students in a class of 30 passing a standardized test than an average teacher, and one of the worst will have three fewer. The impact of teacher quality is much greater than that of school, or even socio-economic factors, but is dwarfed by the variability of achievement within a cohort.

To educate the public about the limited impact that teacher and school quality can have on student achievement, I propose the use of an "index of sensitivity to instruction". If we imagine a test which the lowest scoring eighth-grade students out-perform the highest-scoring seventh-grade students, then we have a test that is highly sensitive to instruction. This would, presumably be a test focusing on content that is taught only to eighth-graders. If we assume the scores to be normally distributed, then a gap of four standard deviations between the seventh-grade mean and the eighth-grade mean would be a reasonable model for this situation. At the other extreme, we can imagine a test that is completely insensitive to instruction; one in which the average score of eighth-graders is equal to that for seventh-graders. Here the gap between the seventh-grade mean and the eighth-grade mean is zero. If we regard these two cases as the ends of a scale, which we anchor at 0 for tests that are completely insensitive to instruction, and 100 for the case where one year's growth equates to four standard deviations, we have a rough scale for measuring sensitivity to instruction[2]. The values of this index for the tests discussed in this paper are shown in table 1.

From the foregoing discussion, it will not be surprising that the tests of achievement in widespread use appear to cluster towards the insensitive end of the scale. NAEP scores 6, TIMSS and the Sequential Tests of Educational Progress (STEP) for middle school mathematics score 8, and even ITBS scores only 10. The 11 tests of mathematics and reading in widespread use in the United Kingdom (Wiliam, 1992) score between 5 and 13 on the sensitivity to instruction scale. While these calculations are inevitably rather "rough and ready", they do show that many, if not the majority, of the tests in widespread use are relatively insensitive to instruction, or at least less sensitive than many people imagine.

[2] Because as noted earlier there is a tendency for the spread of achievement to increase with age, this index is not independent of age.

Test	Sensitivity to instruction score
Completely insensitive test	0
NAEP	6
TIMSS	8
ETS “STEP” tests	8
ITBS	10
Maximally sensitive test	100

Table 1: Values of instructional sensitivity index for selected tests

Conclusion

In this paper, I have argued that the fundamental issue is not that tests are insensitive to instruction; it is that achievement is insensitive to instruction. Put bluntly, most of what happens in classrooms doesn’t change what students know very much, especially when we measure deep, as opposed to surface aspects of a subject. The way we develop assessments, especially when we derive reliability statistics from data generated by a single cohort, exacerbates this, so that we make the tests even more insensitive to the effects of instruction. We can ameliorate the effects of this to some extent, by generating reliability data from a broader sample of students, but this will have only a small effect. If we are to have parents, policy-makers and other stakeholders making appropriate interpretations of school test data, they need to understand that the tests are for the most part, not measuring the things that schools change.

References

Braun, H., Jackson, D. N., & Wiley, D. E. (Eds.). (2001). *The role of constructs in psychological and educational measurement*. Mahwah, NJ: Lawrence Erlbaum Associates.

Brimer, A. (1984). *Wide-span reading test manual* (Revised ed.). Windsor, UK: NFER-Nelson.

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.

Calvert, B. (1958). *Non-verbal test DH*. Slough, UK: NFER.

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

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.

France, N. (1979). *Profile of mathematical skills teacher's manual*. Windsor, UK: NFER-Nelson.

Hanushek, E. A. (2002). *The importance of school quality*. Stanford, CA: Hoover Institution.

Hart, K. M. (1980). *Secondary school children’s understanding of mathematics: a report of the mathematics component of the Concepts in Secondary Mathematics and*

Science programme. London, UK: Chelsea College Centre for Science and Mathematics Education.

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

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.

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

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

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.

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

Quetelet, L. A. J. (1835). *Sur l'homme et le developpement de ses facultés, essai d'une physique sociale [On man, and the development of his faculties, an essay on social physics]*. London, UK: Bossange & Co.

Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1997). *A splintered vision: an investigation of U.S. science and mathematics education*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Shayer, M., & Adey, P. S. (1981). *Towards a science of science teaching: cognitive development and curriculum demand.* London, Uk: Heinemann Educational Books.

Shayer, M., Küchemann, D. A., & Wylam, H. (1976). The distribution of piagetian stages of thinking in British middle and secondary school children. *British Journal of Educational Psychology,* **46**, 164-173.

Shayer, M., & Wylam, H. (1978). The distribution of piagetian stages of thinking in British middle and secondary school children: II – 14- to 16- year olds and sex differentials. *British Journal of Educational Psychology,* **48**, 62-70.

Thorndike, R. L., Hagen, E. P., & France, N. (1986). *Cognitive abilities test administration manual.* Windsor, UK: NFER-Nelson.

Wiliam, D. (1992). Special needs and the distribution of attainment in the national curriculum. *British Journal of Educational Psychology,* **62**, 397-403.

Williamson, G. L., Appelbaum, M., & Epanchin, A. (1991). Longitudinal analysis of academic achievement. *Journal of Educational Measurement,* **28**(1), 61-76.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education,* **11**(1), 57-67.

Yen, W. M. (1986). The choice of scale for educational measurement: an IRT perspective. *Journal of Educational Measurement,* **23**, 299-325.