Measurement, 11: 55–59, 2013

ISSN: 1536-6367 print / 1536-6359 online
DOI: 10.1080/15366367.2013.784165

Routledge
Taylor & Francis Group

How Is Testing Supposed to Improve Schooling? Some Reflections

Dylan Wiliam
Institute of Education, University of London

In "How Is Testing Supposed to Improve Schooling?" Edward Haertel has proposed a framework for thinking about the mechanisms by which testing might improve the various educational processes undertaken in schools. The framework seems to me to be quite general (I use the word "general" here in its mathematical sense of including all cases) since I cannot think of any mechanisms that could not be easily accommodated within the framework.

The distinction between direct and indirect purposes or mechanisms is also useful, although I am not sure that these are the right terms to use. I can see that, from a psychometric standpoint, it may seem that mechanisms that require a test score might seem more direct, whereas the fact that students prepare for a test and, thus, learn more may well be an indirect benefit. However, from a lay perspective, a case could easily be made that the labels should be precisely the other way round. Any benefit that accrues to the test, whether a score is generated or not, could be seen as far more direct than a mechanism that needs to work via a test score. Indeed, from what we know, it seems at least possible, if not likely, that the indirect effects are greater in magnitude. We have known for a while that frequent classroom testing raises student achievement independent of any interpretation made of test scores (see, e.g., Bangert-Drowns, Kulik, & Kulik, 1991), and more recent work suggests that one of the key mechanisms by which testing improves outcomes is by improving retrieval (Little, Bjork, Bjork, & Angelo, 2012). One could go further and point out that some of the effects of testing do not even require a test—merely the possibility of one. If a teacher announces that a spelling test is to take place on the following Friday, provided there is some chance of the test actually taking place and serious consequences attaching to the outcome if it takes place, then the effect on learning is likely to be significant. Note that in this case, it is the speech act of announcing the test, rather than any information arising from the test, that improves the learning. I hasten to add that I do not have a better alternative to the labels "direct" and "indirect," but the fact that these labels may not be immediately meaningful may be worth at least bearing in mind when using these ideas outside the measurement community.

It is also worth pointing out that even where the tests are administered and scored, which would, therefore, presumably count as a direct mechanism, the interpretive argument may be irrelevant. Hanson (1993) offers a useful distinction between a literal and a representative test. To continue the example used by Haertel, the spelling teacher who tests students on the words on a

Correspondence should be addressed to Dylan Wiliam, Institute of Education, University of London, 20 Bedford Way, London, WC1H 0AL, UK. E-mail: dylanwiliam@mac.com

spelling-word list knows that generalizing the results to words that were not tested is unwarranted; the students knew which words were going to be tested and so any increased achievement brought about by students' increased motivation to do well on the test is likely to be limited to the items tested. Of course, the teacher does hope that the results generalize to tests of the same words on other occasions (at least in the not too distant future for example), but this is, nevertheless, what Hanson calls a literal test; the test includes all the outcomes of interest, at least for this week. Of course, most tests are not literal. Most of the time we are interested in test outcomes precisely because of their power to support inferences about things that were not assessed, which is where the interpretive argument comes into play.

For the remainder of this response, however, I want to focus on the first of Haertel's mechanisms—use of assessment for instructional guidance, not least because this is what I have spent the major part of my professional life exploring.

As a result of the work in which Paul Black and I have been engaged over the last two decades, I am now firmly convinced that the use of assessment for instructional guidance offers one of the most powerful ways for improving schooling. However, during this time, I have become increasingly certain that traditional forms of testing have little to contribute here, for three reasons. The first relates to what we are learning about the ability of even well-designed tests to support inferences about multiple aspects of individual performance. The second relates to the "grain size" of the constructs on which the assessment is focused, and the third relates to the practical problems of using such test data in real classroom settings.

DIAGNOSTIC TESTING

It now seems to be generally accepted that most student achievement, at least in the areas in which we are usually interested, is too multidimensional for a single score for each student to be particularly useful. A single score, no matter how carefully derived, cannot do much more than designate a student as either having attained an acceptable level of achievement on some outcome of interest, or having failed to do so. The problem is that there is no "guidance" here, since the assessment outcome does not guide action. It merely confirms that action is needed—the only prescription is that the student needs more and, presumably, better instruction. It seems to me that implicit in the idea of assessment for instructional guidance, therefore, is the idea that the assessment can indicate that some courses of action are likely to be more helpful than others (e.g., that a student would be well advised to spend more time reviewing geometry than algebra) for which a profile of outcomes is necessary. Of course, it would be straightforward to devise a test for each of the profile components, but the testing time required is likely to be prohibitive.

For this reason, much effort in the measurement community over recent decades has focused on the issue of diagnostic testing. The key idea here is that through sophisticated design, the assessment may support inferences about multiple dimensions of performance and produce more reliable component scores than would be possible if the test were made up of a number of subtests, each of which assessed a single component. Many models have been explored for this purpose, including multidimensional item-response modeling, developments of Tatsuoka's "rule space" model and Bayesian inference networks. The problems with these methods is that they are all based on a theoretical analysis of the components of the domain that are presumed to be required for individuals to respond appropriately, not how individuals actually responded. Even

very successful item writers do not appear to be very skilled at writing items that assess multiple skills because, again, what is needed is item writers who understand how test takers actually respond to items (rather than how the item writer thinks they ought to respond). While work in this area is fascinating, and intellectually challenging, I think it highly unlikely that diagnostic testing will provide a substantial improvement in testing efficiency.

THE ISSUE OF "GRAIN SIZE"

The idea that well-designed tests should provide a teacher with information that is useful in planning instruction has obvious attractions. However, what is less obvious is that the "grain size" of the assessment that is most useful for teachers varies considerably from subject to subject. For example, an English language arts teacher may be interested in knowing whether her students can read texts for literal meaning, make inferences that are not explicit in the text, or see how the writer has used particular kinds of language to create a specific effect. This kind of progression may well cover 2 or 3 years' learning for the typical student. On the other hand, the mathematics teacher who plans to teach students to add fractions tomorrow needs to know whether her students can generate sequences of equivalent fractions today. Typically, the grain size of what is needed to plan instruction in mathematics and science is much finer than that in English. One estimate of how many different assessments would be needed to cover the typical content of the middle school mathematics curriculum at the level that would be useful to teachers in planning instruction suggested that around 300 assessments would be necessary (Brown, 1992). For science, the corresponding figure was more than 400. It seems rather unlikely that the time required for so many assessments could be found.

PRACTICAL APPLICATIONS

Even if high quality of students' achievement across 300 different aspects of the mathematics curriculum could be generated, it seems highly unlikely that even the most gifted teachers would be able to use the information effectively. Of course, such models of student achievement could be incorporated into intelligent tutoring systems, but the trade-off between testing time and grain size would seem to be an insurmountable obstacle. Put another way, it seems to me that the kind of precision in testing that the educational measurement community has come to expect is fundamentally at variance with the needs of instruction. In many developed countries, it is accepted that students should be tested for 20 or more hours at the end of compulsory schooling in order to provide sufficiently robust estimates of achievement to support high-stakes inferences. It is unlikely that the same amount of time would be made available for the routine guidance of instruction.

Even if the testing time could be found, there is a second problem with the use of traditional tests to guide instruction and that is that the results often do not arrive in sufficient time to be useful to the teacher. The pace at which instruction occurs in schools means that information that is not available by the day after testing is unlikely to be useful.

Finally, even if the testing time were found and the results were provided in a timely manner, it is far from clear that teachers have the time, or the instructional strategies, to use the information

to improve schooling. Across the United States, many districts have adopted the idea of "data-driven–decision making" which sounds unexceptionable, but if the data are not collected with a clear theory of action about how they are to be used to improve schooling, then little is likely to change. Rather than data-driven–decision making, it seems to me we need a culture of decision-driven data collection—the data are collected only after a clear theory of how they are to be used has been developed, to be certain that they will be usable.

The argument I am making here is that for instructional guidance, teachers simply do not need or find useful (and certainly do not want to wait, or to pay, for) the precision that the educational measurement community is used to providing.

All this may seem like a counsel of despair, so perhaps it is appropriate to conclude these reflections by saying that I am actually very positive about the role that assessment can have in improving schooling.

First, as Haertel points out, often the unit of action is the instructional group rather than the individual student. For this reason, Caroline Wylie and I have been exploring the use of single items that can be embedded in instructional episodes (Wiliam, 2011; Wylie & Wiliam, 2006, 2007). The response of one student to one item is not particularly meaningful, but the response of a class of 30 students to a single item does give the teacher useful information about whether to move on, or to review an instructional episode.

Second, I am convinced that standardized testing has a role to play in raising achievement. It is common for opponents of standardized tests to say that they have no benefit other than for the aspects actually tested. In fact there is a substantial body of evidence that suggests that well-designed high-stakes assessments represent one of the most cost-effective ways of raising student achievement yet devised (Wiliam, 2010).

Finally, one role that assessments can play is less widely appreciated in the United States because of the extensive use of closed tests. Most of the high-stakes tests that students take remain secret at the conclusion of the test. Where items are subject to extensive trialing to minimize bias and other undesirable properties, this is understandable, but it is important to recognize that in other cultures, assessments are routinely made public after their first use. In England, students preparing for school-leaving and university entrance examinations are easily able to obtain every examination paper that has ever been administered for that syllabus. In France, it is not uncommon for questions in the *baccalauréat* examinations used for access to the elite universities to be the subject of discussion in national newspapers the day after the examination. Such scrutiny would of course be uncomfortable at first, but I think would lead to a healthy public engagement in assessment. The resulting alignment of students, teachers, parents, policy makers, and those who design assessments would seem to me to be a force for considerable good in the long term.

REFERENCES

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research*, *85*(2), 89–99.

Brown, M. (Ed.). (1992). *Graduated assessment in mathematics: Complete pack*. Walton-on-Thames, UK: Nelson.

Hanson, F. A. (1993). *Testing testing: Social consequences of the examined life*. Berkeley: University of California Press.

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, *23*(11), 1337–1344.

Wiliam, D. (2010). Standardized testing and school accountability. *Educational Psychologist*, *45*(2), 107–122.
Wiliam, D. (2011). *Embedded formative assessment*. Bloomington, IN: Solution Tree.
Wylie, E. C., & Wiliam, D. (2006, April). *Diagnostic questions: Is there value in just one?* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
Wylie, E. C., & Wiliam, D. (2007, April). *Analyzing diagnostic questions: What makes a student response interpretable?* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.