International comparisons and sensitivity to instruction

Dylan Wiliam
Institute of Education, University of London

Introduction

As one analyzes the idea of international comparisons in detail, the whole venture becomes, to borrow a phrase that Banesh Hoffman used in another context, "quite ridiculously irrational" (Hoffman, 1962, p. 35). Even when issues of population sampling and test administration are dealt with (each of which is a substantial undertaking in itself—see Wiliam, 1998), the issue of the instruments themselves present difficulties that are simply insuperable—not just difficult, but actually impossible. There is no way to take into account the fact that the way that a given word "sits" in one language will not be mirrored in another language (for example the word that is used in the mathematics classroom to describe an "open" set in Chinese would not be used to describe a door that is not closed). However, even as we acknowledge the inherent impossibility of the whole venture of international comparisons, we are left with the inescapable conclusion that, however meaningless the claim, standards of achievement in (say) mathematics really are higher in Japan than Turkey, and almost certainly higher in Finland than in Sweden.

In their fine-grained detail, then, international comparisons are indefensible, but paradoxically, as Barry McGaw's paper demonstrates, they definitely contain some truth in terms of the broad picture—the fact that the rank of most countries changes little whether scores are based on all items, or only those regarded as suitable for each country, is testament to this. It is possible to argue, therefore, that as the major uses made of such international comparisons tends to be at the coarsest level, that little harm is done. However, I think that while such a view has some empirical support, I believe that the conclusion is unsound, because of a little-discussed, but highly significant aspect of education assessments—their sensitivity to instruction.

Data from international comparisons are used for many purposes, but most of the uses, and certainly the uses highlighted by the OECD, appear to be focused on the outcomes of the comparisons as indicators of the quality of educational provision in the countries involved. The implication is that we can look at those systems that appear to be most successful, and possibly derive some ideas for improving systems that are less successful. The central assumption here is that these assessments are valid measures the effectiveness of the system. This claim involves two aspects. The first is that the assessments used do indeed measure valued outcomes, and the second is that differences observed are the result of differences in the quality of educational provision. The first of these is more or less taken care of by the item and test development processes that McGaw describes, but the argument I want to propose in this paper is that these processes of item and test development actually weaken the ability of the international comparisons to provide evidence about the quality of educational provision.

Of course, it could be argued that the differences in the performance of different countries are, by definition, the result of differences in the quality of educational provision, if only we define the term "educational provision" broadly enough. However, it is my contention that the intended inferences are generally more focused that this, and specifically that differences in country scores are the result of differences in the quality of *instruction*. The important question, therefore, is to what extent are the assessments used in international comparisons sensitive to instruction?

Sensitivity to instruction

If educational assessments are to be used to draw inferences about the quality of instruction, then it seems obvious that we need assessments on which students who have been exposed to high quality instruction do well, and students who have been exposed to low quality instruction do much less well. However, it is far from clear that this is the case with the tests used for international comparisons.

The first point to make here is that learning itself is relatively insensitive to instruction in that the progress made by individual students is rather slow compared to the variability within the cohort. For example, Rodriguez (2004) found that on the tests used for TIMSS, the difference between seventh-grade students and eighth-grade students in the US was around one-third of a standard deviation. In other words (assuming a range of two standard deviations above and below the mean) the progress made by an age-cohort in one year is one twelfth of the range of achievement within the cohort. Or to put it another way, within the cohort of students of seventh-grade age, some will be performing like average first graders, and others at a higher level than the average twelfth-grader. Wiliam (2007) analyzed a number of tests, both standardized and unstandardized, from the UK and the USA and found that for most tests, the annual increase in performance ranged from 0.25 to 0.4 standard deviations.

What this means in practice is that the impact of high-quality instruction on student performance is less than appears to be generally assumed. 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 (Wright, Horn & Sanders, 1997), but is dwarfed by the variability of achievement within a cohort.

The second point is that the procedures used in almost all test construction decrease the sensitivity of the test to instruction. 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.

The third point is that the specific procedures used to develop tests for international comparisons decrease the instructional sensitivity yet further. Jones (1993) reported an interesting example of differential item functioning (Holland & Wainer, 1993) in which an item on similar triangles was answered much more successfully by students who took the item in a Welsh-medium test than those who took the same item in the medium of English, even when overall achievement on the test was taken into account. The reason for this is that the word "similar" is used in English both in general conversation where it means "having a resemblance" and in the secondary school mathematics classroom, where it means that two shapes have the same angles, and the corresponding sides are proportional. In Welsh however, the word used in the mathematics classroom, "*cyflun*", is rarely, if ever used outside the classroom. A question that asks students whether two shapes are similar or not is likely to cue a mathematical response when posed in Welsh to Welsh-speaking students, while the English-medium version is more likely to cue students to look for general resemblances between the shapes. The procedure that McGaw describes, in which items that exhibit differential item functioning between different languages are omitted, would require the removal of items that assess the concept of "similar triangles". Through this process items with strong links to instruction are likely to be systematically removed, thus decreasing the sensitivity of the test to instruction further.

Another example of this phenomenon occurs in testing secondary school physics. One of the crucial distinctions that arises in kinematics is the distinction between speed (a scalar quantity) and velocity (a vector quantity). In English, the non-technical term, speed, denotes the scalar quantity while the technical term, velocity, denotes the vector. In Welsh, these two are reversed. The everyday term for how fast something is going is *cyflymder*, but in the science classroom, this term denotes velocity, not speed (the reason for this is that in Welsh, acceleration—a vector quantity—is *cyflymdra*, so it was decided that it made more sense for the vernacular term, *cyflymder*, also to denote a vector quantity). A new term, *buanedd* (literally, "soon-ness") was coined for the scalar quantity. So, in an item that uses

the word speed, candidates taking the item in Welsh are likely to be at an advantage, while an item that uses the word velocity will advantage candidates taking the item in English

A common method for ensuring that translation of items are correct is through a process known as "back-translation". The idea is that an item is translated from the source language to the target language, and then the translation is independently translated from the target language back to the source language. If the item resulting from this dual translation is substantially the same as the source item, then this provides evidence of the fidelity of the original translation from the source language to the target language. Such a process does identify the more egregious examples of poor translation, such as the probably apocryphal story of a machine translation from English to Russian, and then from Russian to English of the phrase "Out of sight, out of mind" which produced, after back-translation, the phrase "invisible idiot". However, such back translation would not identify the problem of speed and velocity in English and Welsh discussed above, because, in terms of the classification proposed by Poortinga (1995), we have *identity of concepts* between English and Welsh; in terms of their meanings, *cyflymder* is precisely equivalent to velocity and *buanedd* is precisely equivalent to speed. What differs is the way that these words "sit" in the language. They do not differ in what they denote, but they differ in what they connote. An item that is couched in terms of speed would be likely to show DIF in favour of students taking the test in the Welsh-medium, and an item couched in terms of velocity would be likely to show DIF in favour of the students taking the test in the English-medium. Both would be rejected by the procedures used in PISA.

Conclusion

In this paper, I have argued that, in broad terms, differences in scores on international comparisons undoubtedly point to robust differences in the quality of educational provision. However, as we dig deeper, we find that the procedures of test construction, and specifically the development of items in multiple languages, decrease the sensitivity of the tests to instruction in ways that are not fully understood, and which may vary in important ways from language to language. As long as differences in test scores are interpreted at the broadest level, this may not matter too much, but as more and more sophisticated analyses are conducted, it becomes harder and harder to interpret these differences. The differences between countries in the proportion of variance in student scores attributable to within school effects, between school effects attributable to social backgrounds of schools, of students, and between school effects not attributable to social backgrounds, are of great interest, but interpreting them is difficult. In particular, it is difficult to discount the possibility that some of these differences may be caused by the fact that the tests in different languages may be differentially sensitive to instruction.

References

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

Hoffman, B. (1962). *The tyranny of testing*. New York, NY: Crowell-Collier Press.

Holland, P., & Wainer, H. (Eds.). (1993). *Differential item functioning: theory and practice*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Jones, D. (1993). Words with a similar meaning. *Mathematics teaching*(145), 14-15.

Poortinga, Y. (1995). Use of tests across cultures. In T. Oakland & R. K. Hambleton (Eds.), *International perspectives on academic assessment* (pp. 187-206). Boston, MA: Kluwer Academic Publishers.

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

Wiliam, D. (1998). Making international comparisons: the Third International Mathematics and Science Study. *British Journal of Curriculum and Assessment,* **8**(3), 37-42.

Wiliam, D. (2007). *An index of sensitivity to instruction.* Paper presented at the Symposium entitled "Three practical policy-focused procedures for determining an accountability test's instructional sensitivity" at the annual conference of the American Educational Research Association held at Chicago, IL.

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.