ASSESSMENT: MORE AN ART THAN A SCIENCE

John Maratos

Vlaardingerbroek (1995) offers a way around obstacles to grade comparability when different schools use different tests. He notes that the increasing reliance on internal, school-based assessment requires attention so as not to impede comparison. The solution he proposes is on its face sound. His reasoning, however, is questionable.

The solution proposed involves internal, school-based assessment, moderated by an external common assessment with both, as far as practicable, weighted equally. Vlaardingerbroek implies that it is not fool-proof, however. Teachers tend to object to norm-referenced testing. They tend to prefer internal skills-based assessment; even for subjects unsuited to a break-down into a set of necessary and sufficient component skills, or techniques. This may hinder progress. A related danger not mentioned is that even subjects thought to be clear candidates for skills-based assessment, such as mathematics, can suffer a loss of coherence when they are also taught as mere sets of skills.

Nonetheless, the proposed solution is sound up to a point. It is arrived at by a route that creates unnecessary anxiety about comparability, largely because it neglects two central and related issues: (i) the limits of statistical/empirical research in informing decisions about assessment procedures; and (ii) the danger of inexplicitly conflating statistical and non-statistical reasoning. The conceptual route to the proposed solution needs refining.

Two measures by which to secure common standards and assessment comparability might be: (i) enforce uniformity in difficulty levels and skills tested across schools, thus initiating similar external tests; or (ii) permit variation in difficulty levels and skills, but require some common external assessment - which schools could organise - and make all results publicly available.

The first measure is extreme. It could eventually stifle intellectual development by halting the ebb and flow in curricular responses to general and enduring ideas and information, as well as to local and timely concerns in schooling. Curricula would harden and become brittle. Rather than strategic thinking on such issues as student failure, reactionary responses would dominate.

The second measure permits vitality, by leaving unfettered the contact between the educational concerns of students, parents, teachers, academics, politicians and so on. As long as schools retain a say in the content and the extent of common, external assessments, the ensuing tension will not necessarily prohibit diversity, nor comparability. Diversity can be contained in the long run as a function of publicity. Norm-referencing will operate as schools' self-imposed curricular discipline.

Containing diversity, however, does not guarantee comparability. To dispel this fear, let us recall that school assessment need not slavishly serve a single end. It can serve a variety of desirable ends, which include comparability. Indeed, significant education demands that school assessment address more than the need for comparability, pure and simple. Diverse curricula present no necessary obstacle to viable inter-school test comparisons, as long as the diversity occurs around a central cluster of shared concepts, difficulty levels and skills. Local circumstances will influence the speed and certainty with which it would happen, but even schools strongly tied to external examinations can have some authority gradually devolved to them. It is where authority and public scrutiny have been diffuse, weak and optional that success might prove extremely elusive.

There is an ever-present pressure for the acquisition of concepts, skills and talents to serve this world. Ultimately, many administrators, teachers and parents will resist the drift of curricula into incoherence. For instance, ample information and tests exist outside schools on what constitutes accurate addition, subtraction, location of rivers, naming of cities, significant historical dates and so on. It is also difficult to conceal forever such things as who the best cancer doctors are, the safest airline companies, the most reliable civil engineers and so on. Public resistance to curricular chaos, assisted by publicity of school assessment results, improves the probability of genuine and comparable educational progress in diverse schools.

For although comparability is certainly crucial for an informative mark, it is never the only concern dealt with by assessment. Statistical decisions related to the type of external moderation used are actually embedded in a larger body of decisions about assessment, many of which are non-statistical. The limits of statistics in developing assessment procedures can be stated simply as that assessment is an art informed by science. It is not an emerging science still struggling to divest itself of some primitive normative interests. The point and structure of examinations, whether criterion or norm-referenced, cannot be fully accounted for by a mere explication of constitutive techniques.

The second issue is in effect the compounding of the defects described above. Neglect of the appropriate place of statistical reasoning in developing assessments, is necessarily bound up with conflating non-statistical and statistical inferences. Rather than spelling this out explicitly, Vlaardingerbroek's (1995) discussion gives the impression that calculators and computers will suffice for all the decisions at hand.

For instance, at the very start of his article he complains that percentages ascribed to examination results may orient the illusion of being informative on a general level, across schools. And yet, school programmes may be radically dissimilar. But the complaint is not an essentially statistical one: it is in fact ethical, implying a plea of justice: give students their due by making their test scores as precise and comparable as possible. Only then can parents, students, administrators and employers know what schooling really has achieved. Leaving these considerations inexplicit helps create its own illusion: namely, that the adjustments to weightings of the examined components described later are advisable on purely statistical grounds. The decision that led to a statistically significant decline in weightings of examinations in New Zealand since 1985, for example, is portrayed as a statistical inference from statistical premises. So, too, is the following situation:

> All other factors equal, it was calculated that a school which places a 25% weighting on IA (internally assessed) instruments arising from laboratory work, can expect its students to perform 3.0% better on average in the external examination than a school that places no IA weighting on lab work at all. The research revealed the fallacy inherent in approaches that use IA merely as a series of trial runs for the exam. (p.10)

No doubt the research did show all this. But why imply that all assessment decisions are statistical? The decision to increase the weighting of lab work is not statistical/empirical, but logical. Good teaching demands that what is internal to a subject be assessed according to its constitutive importance. Masters of the subject who have thought through its teaching decide what that weighting will be. Statistics inform the numerical representation of these prior logical decisions for the purposes of constructing tests, marking and grading them, even comparing them.

The situation is not helped by talk of "a positive correlation between the weighting of practical work in schools' assessment schedules and their students' mean performance in an external examination" (p.10). The spell of science, of the naive faith in hard empirical verification alone, encourages the belief that there is incontrovertible, objective evidence to make corresponding adjustments to curricula. Increase lab work and more students will succeed in science. Make science more hands on! But unless other relevant variables are identified and controlled, this pseudo-statistical inference may lead to disaster. These ostensible positive correlations - note the treacherous ambiguity, not entirely unexploited: 'positive' meaning high degree of regular coincidence and 'positive' as a practical imperative - tell us nothing about the standard of the lab work being introduced. They tell us nothing about how well science was taught before lab work was assessed more thoroughly, and how well afterwards. They tell us nothing whatsoever about a durable approach to assessment, one grounded in considerations internal to students' educational interests. Even when misread as a practical imperative, the above statistics merely imply that there is an instrumental connection between high marks in external science exams and greater IA weighting on lab work. Statistics will also show a positive correlation between seeing examination questions prior to the examination and scoring high marks. Why not also exploit this very direct instrumental connection?

Vlaardingerbroek is right to imply that norm-referencing is unavoidable and that assessment design is best approached with both eyes open. His preferred solution is also sound. But his

reasoning suffers from opacity. There are unavoidable logical and ethical decisions behind statistical and other empirical research on assessment. Keeping them out in the open allows scrutiny and assists in their improvement. Statistics have a critical role to play in assessment decisions, but they should be kept in their proper place. Educationists must avoid giving the impression of torturing statistics in the hope of extracting from them the secrets of an educational philosophy.

Reference

Vlaardingerbroek, B. 1995. Assessment: More Than Just Marks, *Pacific Curriculum Network*, 4 (1): 8-11.