Educational Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713699076

English national curriculum assessment: a commentary from the USA - or exhibiting kindness to the colonies

W. James Popham [a]
[a] University of California, Los Angeles, California, USA

To cite this Article Popham, W. James(2009) 'English national curriculum assessment: a commentary from the USA - or exhibiting kindness to the colonies', Educational Research, 51: 2, 265 — 268
To link to this Article: DOI: 10.1080/00131880902891966
URL: http://dx.doi.org/10.1080/00131880902891966

Educational Research
Vol. 51, No. 2, June 2009, 265–268

Routledge
Taylor & Francis Group

COMMENTARY

English national curriculum assessment: a commentary from the USA – or exhibiting kindness to the colonies

W. James Popham*

University of California, Los Angeles, California, USA

Because the US Constitution does not identify education as a federal responsibility, provision of public schooling falls to America's states. Such is the case with any governmental obligation not specifically staked out in the Constitution as a federal function. Although differences exist, sometimes fairly substantial ones, in how our 50 states assign educational duties to their local school districts, Americans have historically regarded education as a 'states rights' arena – not a federal activity. The same perception applies to the assessment of students. Educational testing has traditionally been regarded in the US as a state rather than federal undertaking.

Oh, it is true that in 1969 we installed the National Assessment of Educational Progress (NAEP), but this is a set of sampling-based administered periodically with no consequences tied to students' performances. NAEP is a no-stakes test. Thus, if you could time-travel back a few decades or so and ask almost any American educator whether federal testing of the nation's students is appropriate, the answer you would surely get would be, 'Of course not – that's a state responsibility.'

But this states rights view of educational testing has, in recent years, been softening. Perhaps because of increasing dissatisfaction with the quality of US tax-supported schools, we now find editorial essays, sometimes written by prominent US educators, arguing for the installation of national, federally administered educational accountability tests. This departure from what, in former years, was an almost sacrosanct view about who should control educational testing represents a profound shift in the thinking of many US policymakers.

Interestingly, most of the measurement quandaries faced at the national level in England since 1988 have proved equally vexing to state-level assessment officials in the US during recent decades. In the 1970s, for example, there was a strong push in most US states to devise 'minimum competency' tests that, when administered to a state's students, would demonstrate those students could display at least a rudimentary level of basic-skills mastery. Yet, given the often laughably low levels of skills ostensibly measured by these tests, it was impossible for the same tests to accomplish simultaneously another alleged purpose, namely, to focus educators' attention on praiseworthy curricular aspirations. So, in a manner remarkably similar

*Email: wpopham@ucla.edu

ISSN 0013-1881 print/ISSN 1469-5847 online

DOI: 10.1080/00131880902891966
http://www.informaworld.com

to the 'multiple-missions' issue faced by England's assessment specialists, their counterparts in America's state assessment programmes have wrestled with the issue of how many measurement functions a high-stakes tests can legitimately serve.

Against a shifting set of assessment preferences in the US regarding whether educational assessment should continue to be a states rights game or become a federally dominated undertaking, the publication of five first-rate analyses about England's national curriculum assessment (NCA) is particularly propitious. Taken together, these five papers (Whetton 2009, this issue; Stobart 2009, this issue; Newton 2009, this issue; Wyse and Torrance 2009, this issue; Green and Oates 2009) in this special issue of *Educational Research* provide an insightful collection of views regarding the perils and payoffs of national educational assessments.

Although I thought the five papers, separately and in concert, were terrific, I am among those who view history's and other nations' 'lessons' with great wariness. An assessment approach that might have worked wonderfully in England at a particular moment in time, and under the direction of an adroit administrator, might prove disastrous in the US if administered at a different time by a less adept leader. NCA 'lessons' dare not be transplanted unthinkingly to other nations.

To illustrate the dangers of simply lifting procedures or concepts from one setting and plopping them into another, consider the conception of 'criterion-referenced' measurement that seems to be present in certain quarters of England's assessment thinking at the outset of the national assessment. When Robert Glaser first drew our attention to this way of interpreting test-takers' scores way back in 1963 (Glaser 1963), in contrast to a more traditional 'referencing' of an examinee's score back to the performance of a norm group or previous test-takers, he actually presented *two* conceptions of what was meant by a 'criterion.' A criterion was described by Glaser, first, as a *level* of performance and, second, as a *criterion behaviour*, i.e. as a well described skill or body of knowledge. Although for a number of years in the US, these two meanings of 'criterion' were the source of considerable confusion among members of America's measurement community (Popham and Husek 1969), now almost all US assessment specialists who work with criterion-referenced assessments regard a test's criterion as a clearly explicated set of skills or knowledge – not as a level of performance. Yet, if US assessment authorities unthinkingly adopted England's level-oriented conception of a criterion-referenced testing, we would be choosing a measurement strategy that is fundamentally different from the way most US states using criterion-referenced tests develop these kinds of assessments. Nonetheless, if one remains alert to the possibility of such differences, the five papers about NCA are simply brimming with thought-provoking ideas over which some serious mulling is surely warranted. Let me say a few words about each paper.

Whetton (2009, this issue), opens with a brief history of NCA in England from 1989 to the present. He tells readers a good story, lacing it with recountings of real-world people and events to make one want to see 'what happens next.' I suppose, if obliged to choose between Whetton's succinct history of NCA and a cracking good espionage novel, I'd probably opt for the latter. But his is a remarkably readable depiction of events that otherwise might have seemed sublimely stuffy. This rapid-fire account of England's often politicized dance with national testing presents a series of choice points that, though made one way by English policymakers, could clearly be made in other ways elsewhere.

In Stobart's (2009, this issue) pithy tour of assessment validity, readers will find an especially solid 'nutshell' description of what is meant by validity in an

educational measurement context. He comes down clearly in favour of a conception of validity in which the users of a test must appraise that instrument not only on the accuracy with which it measures whatever it is measuring but also, and this is a crucial distinction, the defensibility of the *uses* to which the test's results are put. Validity, according to Stobart, lies not only in the accuracy of the inferences about the meaning of a test-taker's score, but also whether the results of a test are put to proper use. This notion is fundamental to the concerns registered by Stobart and other authors about the multiple uses England's curriculum assessments are currently asked to satisfy.

Reliability, along with validity, one of the two meat-and-potatoes staples of educational testing, is given thoughtful scrutiny by Newton (2009, this issue). I was particularly impressed with his concluding concerns about the wisdom of relying on Cronbach's Coefficient alpha as the dominant determiner of reliability for England's curriculum tests. Coefficient alpha, along with a number of similar reliability indices, are measures of a test's 'internal consistency', i.e. the degree to which a test's items are functioning in a homogeneous manner. Well, if one of the purposes of a national curriculum test is to help teachers make more sensible *instructional* decisions, then a test in which all of the items are functioning homogeneously cannot, by definition, be diagnostic. As Newton implies, although England's national tests may be striving for reliability (after all, who would opt for unreliability?), it is likely – at least for NCA's promotion of instructional improvement – that the wrong index of reliability is being employed.

Because I believe the most important mission of all educational tests is to improve instructional quality, and educational tests that fail to improve students' learning should be summarily shredded, I was particularly interested in the Wyse and Torrance paper (2009, this issue). Their conclusions are both sobering and saddening, namely, that England's bold experiment in using tests to benefit learners is not only narrowing the curriculum but also driving teaching toward instructional ineffectiveness. These authors call for a renewed focus on formative assessment, surely to be used in a manner consonant with the considerable empirical evidence now at hand regarding how the formative assessment process can benefit students. Any national assessment programme that diminishes the use of assessment to help students learn better is a programme that should not exist. Wyse and Torrance make this point powerfully.

Given the need to satisfy multiple measurement missions, Green and Oates (2009, this issue) provide readers with three models for simultaneously improving learning, supplying school-accountability information, and providing evidence on system performance for policymakers. These authors believe that 'the use of national test data for multiple purposes has a negative impact.' Nonetheless, Green and Oates do a commendable job or setting forth meaningfully different strategies for accomplishing these three missions.

While there is certainly evidence in several papers that the pursuit of multiple purposes by one test can be educationally detrimental, it is not clear whether the tests being considered in these analyses were deliberately constructed with specific purposes in mind. If you want a test to serve as a catalyst for improved teaching, you incorporate certain features in that test. If you want the test to hold teachers accountable, you build in other features. Yes, you can even accomplish more than one function with a given test if you deliberately design the test to tackle two or more missions. Yet, the question of *how* England's national tests were constructed and the

precise measurement functions, if any, those tests were intended to accomplish is not addressed in any depth by the authors of these five papers. It is most definitely not true that, 'a test is a test is a test.' I am certain the authors of these five papers know this all too well. But in analyses such as these it must be emphasized and re-emphasized that the *nature of the test* determines whether a test can serve more than one master.

As noted at the outset, increasing numbers of policymakers in the US are now beginning to call for national rather than state tests. If those advocates of national tests ever get the upper hand, then I hope the publishers of *Educational Research* will instantly bundle up large numbers of these five excellent papers and fire them off to American policymakers. After all, to enlighten assessment thinking in this former colony of Great Britain would clearly be the proper thing for the Motherland to do.

References

Glaser, R. 1963. Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist* 18: 519–21.

Green, S., and T. Oates. 2009. Considering alternatives to national assessment arrangements in England: possibilities and opportunities. *Educational Research* 51, no. 2: 229–45.

Newton, P. 2009. The reliability of results from national curriculum testing in England. *Educational Research* 51, no. 2: 181–212.

Popham, W.J., and T.R. Husek. 1969. Implications of criterion-referenced measurement. *Journal of Educational Measurement* 6: 1–9.

Stobart, G. 2009. Determining validity in national curriculum assessment. *Educational Research* 51, no. 2: 161–79.

Whetton, C. 2009. A brief history of a testing time: National curriculum assessment in England 1989–2008. *Educational Research* 51, no. 2: 137–59.

Wyse, D., and H. Torrance. 2009. The development and consequences of national curriculum assessment for primary education in England. *Educational Research* 51, no. 2: 213–28.