Book reviews

Book review: a language tester looks at *The bell curve*

Herrnstein, R.J. and **Murray, C.** 1994: *The bell curve: intelligence and class structure in American life.* New York: The Free Press. 845 pp.

In *The bell curve: intelligence and class structure in American life*, Richard J. Herrnstein and Charles Murray have addressed a complex topic: the role of cognitive ability in USA society. The book is devoted to a thesis which the authors summarize on p. 528: 'Our central concern since we began writing this book is how people might live together harmoniously despite fundamental individual differences.' Theirs is a record of evidence of those 'fundamental individual differences' in cognitive ability and of their social cost, and they close the book with social and cultural suggestions to accommodate such differences.

Their book is anchored in the USA context. *Language Testing* is an international journal and, at the surface, it may seem that a review of *The bell curve* is inappropriate for inclusion here. However, second/foreign language assessment – perhaps because of its multinational academic personality – is in a unique position to evaluate the central argument of *The bell curve*, a point to which this review will return.

The substance of Herrnstein and Murray's argument is divided into four general parts comprising 22 chapters. The four parts are labelled '(I) The emergence of a cognitive élite', '(II) Cognitive classes and social behavior', '(III) The national context' and '(IV) Living together'. There is an introduction and various front matter, and there are seven appendices, which amplify technical considerations of the book's analyses. The book is heavily footnoted for bibliography and content, and the content footnotes provide amplification of many topics and issues.

The bulk of their evidence for the presence for sizeable social effects of cognitive ability comes from a large database, the National Longitudinal Survey of Youth (NLSY): '. . . the NLSY is a very large (originally 12,686 persons), nationally representative sample of American youths who were aged 14 to 22 in 1979, when the study began, and have been followed ever since' (p. 36). Herrnstein and Murray selected the NLSY over other national databases because

> Only the NLSY combined detailed information on the childhood environment *and* parental socioeconomic status *and* subsequent educational and occupational achievement *and* work history *and* family formation *and* –

> crucially for our interests – detailed psychometric measures of cognitive skills (p. 119, emphasis in original).

The cognitive psychometric measures which they use came from a measure used in the NLSY:

> . . .the Armed Forces Qualification Test, the AFQT. It is what the psychometricians call 'highly g-loaded', meaning that it is a good measure of cognitive ability . . . Because the raw scores on the AFQT mean nothing to the average reader, we express them in the IQ [Intelligence Quotient] metric (with a mean of 100 and a standard deviation of 15) or in centiles. Also we will subsequently refer to them as 'IQ scores,' in keeping with our policy of using IQ as a generic term for intelligence test scores (p. 120).

That is, Herrnstein and Murray defined their dependent variables as measuring a unidimensional cognitive ability trait which they chose to transform to the IQ metric. This reflects the essential statistical epistemology of the book.

Each of the four parts serves a different function. Part I builds a case for the evolution of a 'cognitive élite' in USA society. They argue that this élite class, basically a phenomenon of this century, is in danger of evolving into a unique, new ruling class. That latter point is elucidated further in Chapter 21 in Part IV. Parts I and IV work together to raise a spectre (the evolution of a cognitive élite) and discuss how social efforts to undo that have failed and that other social and cultural changes are needed.

In Parts II and III the authors argue that intelligence is measurable and its social role noteworthy. Cognitive ability, as operationalized by the AFQT, is treated both as a dependent and an independent variable. In Part II, Herrnstein and Murray argue that it covaries with many phenomena, such as poverty, schooling, idleness and welfare. During their exposition in Part II, they present their understanding of the literature on the subject that IQ is between 40 and 80% inherited and, throughout the book, they use the 'middling estimate' (p. 105) of 60%. In Part III, they explore some possible sources for variation in cognitive ability: ethnicity, demography and the '. . . social behavior and the prevalence of low cognitive ability' (p. viii). In the introduction to Part III they state:

> Nothing seems more fearsome to many commentators than the possibility that ethnic and race differences have any genetic component at all. This belief is a fundamental error. Even if the differences between races were entirely genetic (which they surely are not), it should make no practical difference how individuals deal with each other. The real danger is that the élite wisdom on ethnic differences – that such differences cannot exist – will shift to opposite and equally unjustified extremes. Open and informed discussion is the one certain way to protect society from the dangers of one extreme view or the other (p. 270).

In Part IV, they close with proposals for social change that they argue will help avoid the evolution of a detrimental cognitive élite. The final chapter is entitled 'A place for everyone' and in it they appeal to the civic tradition of the USA:

> We are asking you to consider . . . that the Founders were fully aware of how unequal people are, that they did not try to explain away natural inequalities, and that they nonetheless thought the best way for people to live together was under a system of equal rights (p. 530).

To effect this, Herrnstein and Murray propose that each person be accorded a 'valued place in society' (p. 535), and they note: '. . . most people by far have enough intelligence for getting on with the business of life . . . Our solutions assume that the average American is an asset, not part of the problem' (p. 536). They follow with precise suggestions to do this, e.g., restoration of community social functions, simplification of social rules, reducing credentialism (e.g., the awarding of jobs based solely on possession of a college degree), simplification of punishment for crime and '. . . restoring marriage to its formerly unique legal status' (p. 545) by disallowing benefits to unmarried couples. They also state that child-bearing policies '. . . represented by the extensive network of cash and services for low-income women who have babies, be ended' (p. 548).

It is not the intent of this review to dwell on the merits of the precise social and cultural suggestions which Herrnstein and Murray make – to do so is in the territory of political science, philosophy and sociology and not the territory of *Language Testing*. Rather, this review examines *The bell curve* from the standpoint of human mental measurement informed by second/foreign language education. This is a *measurement* review of *The bell curve* and not a *social review*.

Herrnstein and Murray have built their entire case on a single epistemology: statistically developed norm-referenced tests. Norm-referenced measurement (NRM) is a test development technology yielding scores which distribute themselves normally so that results are interpretable as ranks; it is a refined technology that is about one hundred years old. The authors know the NRM trade. It is clear that they are statistically sophisticated and can speak the psychometric language of NRM quite fluently. Their technical acumen is very impressive.

What is absent is that to which language testers have learnt to attend most closely: the content of the test. Second/foreign language assessment is uncomfortable with tests for which items and tasks are developed based solely or largely on their statistical characteristics. This discomfort is not new. For example, the extensive use of nonstatistical descriptive rubrics in the USA government began just after the second world war, as the Foreign Service Institute scale was born. Later, in 1980, Canale and Swain published their watershed article which argued elegantly for at least three major mental traits which combine into communicative competence. This argument was picked up, investigated empirically and reasoned further by later scholars (see Bachman and Palmer, 1982; Canale, 1983; Bachman, 1990: Chap. 4). As a result of this attention to content, language testing has also become a proving ground for criterion-referenced measurement (CRM), an epistemology that prioritizes content over statistics in test development (see Davidson and Lynch, 1993; Lynch and Davidson, 1994). Herrnstein and Murray are quite the opposite – to the extent that they

attend to content (if at all), it is subjugated to statistics. Their praise of a 'highly g-loaded' measure, noted above, is a prime example, as is their treatment of content of NRM IQ tests, which is at best anecdotal (e.g., the digit span example on p. 283).

Their embrace of statistics is very tight. They seem to have ignored vast literature on the nature of measurement validity, preferring, instead, to define validity solely as predictive power. This limited vision of validity is clear throughout the book, and it is stated overtly in footnote 1 to Appendix 5 on p. 770:

> Validity is measured by the correlation between predictor and outcome, which, multiplied by the ratio of the standard deviations [*sic*] of the outcome to the predictor, gives the regression coefficient of the outcome on the predictor.

To define validity solely as a predictive correlation is wrong on two counts. First, it ignores the well established alternate validities: construct, concurrent and content. Secondly, and perhaps more importantly, it ignores more recent literature on a unified vision of validity, where construct evidence, predictive/concurrent evidence and content evidence must work together to form a reasoned argument for the truth of a test. (For a good overview of recent unified approaches to validity, see Gronlund, 1993: Chap. 10; for some original literature on this matter, see Messick, 1989, or Shepard, 1993.) It is not surprising that a book which draws so heavily upon norm-referencing would, of necessity, define validity as predictive power.

The authors are quite open about their worldview. At the beginning of the book, they clearly admit their epistemological leaning by acknowledging the '... assumption that intelligence is a reasonably well-understood construct, measured with accuracy and fairness by any number of standardized tests' (p. 1). Regardless of whether or not that assumption is correct, the book's reasoning follows from the implacable warrant that cognitive ability exists, that it is largely a single general trait and that any number of NRM tests exist which measure it adequately, chief among them those used in the NLSY. The reader must buy into that assumption to buy the rest of what the book has to say.

The readership of *Language Testing* comes from many 'measurement cultures' (Davidson and Bachman, 1990). This readership decides how to assess language ability in many ways, all conditioned by varying mandates and realities which cross national boundaries. It is a unique entity. But one common theme crosses those boundaries: regardless of the cultural demands, language testers all measure the same thing – language. Attention to content and a broader view of validity are, as they should be, implicit in the development of language tests.

Possibly, Herrnstein and Murray could have added extensive discussion of a model of cognitive ability – a description of the nature of the mind. Possibly, they could have built a CRM case as well as an NRM case on how to measure that model. Possibly, they could have done so and still made some of the provocative social points they did make. But they did not do so. Instead, they rested their entire argument on a technically complex but

epistemologically simplistic statistical worldview. As a result, their argument will doubtless remain unconvincing to the readers of this journal, who take a broad conception of their charge. Language testing is building a stronger argument on the nature and impact of its trait than Herrnstein and Murray have done for theirs.

References

Bachman, L.F. 1990: *Fundamental considerations in language testing.* Oxford University Press.

Bachman, L.F. and **Palmer, A.S.** 1982: The construct validation of some components of communicative proficiency. *TESOL Quarterly* 16, 449–66.

Canale, M. 1983: On some dimensions of language proficiency. In Oller, J.W. jr, editor, *Issues in language testing research.* Rowley, MA: Newbury House, 333–42.

Canale, M. and **Swain, M.** 1980: Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1, 1–47.

Davidson, F. and **Bachman, L.** 1990: The Cambridge-TOEFL comparability study: an example of the cross-national comparison of language tests. In de Jong, J.H.A.L., editor, *Standardization in language testing: AILA Review* 7, 24–45.

Davidson, F. and **Lynch, B.** 1993: Criterion-referenced language test development: a prolegomenon. In Huhta, A., Sajavaara, K. and Takala, S., editors, *Language testing: new openings.* Jyväskylä, Finland: University of Jyväsklyä, 73–89.

Gronlund, N.E. 1993: *How to make achievement tests and assessments.* Boston, MA: Allyn & Bacon.

Lynch, B. and **Davidson, F.** 1994: Criterion-referenced language test development: linking curricula, teachers and tests. *TESOL Quarterly* 28, 727–43.

Messick, S. 1989: Validity. In Linn, R.L., editor, *Educational measurement* (3rd edn). New York: NCME/ACE–Macmillan, 13–103.

Shepard, L.A. 1993: Evaluating test validity. In Darling-Hammond, L., editor, *Review of research in education 19.* Washington, DC: AREA, 405–50.

Fred Davidson
University of Illinois at Urbana Champaign