Introduction

HARVEY GOLDSTEIN & ANTHONY HEATH

EDUCATIONAL STANDARDS HAVE FIGURED PROMINENTLY in recent debates over educational policy in the UK and elsewhere. Despite this, or even perhaps because of it, there is little clarity about the nature of a 'standard', little understanding of how such debates are situated historically, and scant awareness of measurement issues. The British Academy invited a number of distinguished academics and researchers to present papers at a one day symposium designed to bring together a number of perspectives on this issue. This symposium was held on 9 October 1998 and a list of participants is given in the appendix. The speakers and discussants were chosen for their expertise in a number of relevant areas, together with an audience which contained other academics and researchers as well as policy makers.

None of the contributors claims to offer a straightforward 'solution' to the problems of definition and measurement, or to be able to provide prescriptions for official policy. Rather, they have attempted to provide an analysis of the nature of the problems and a contextualisation of the debates, both historically and cross-nationally. In this way we hope better to inform public debate. We believe that this is the first serious attempt to bring together such a distinguished collection of scholars on this topic, at least in the UK, and we would recommend these contributions to academics and policy makers alike.

In the first section of the volume Alison Wolf addresses the issue of how far there is an international consensus about the way in which educational standards should operate. She points to important differences between countries. In the USA, for example, the SAT and GRE have become enormously important high-stakes tests for young people

Proceedings of the British Academy, **102**, 1–8.

seeking entry to higher education courses. They are standardised tests in the sense of machine-readable multiple-choice items, selected on the basis of psychometric criteria. Judgements about content and item format, and therefore what a given level of success actually involves or means are buried from sight. The 'scientific' basis of the test construction seems to be associated with a high level of public confidence in the objectivity of the tests. But unlike the case in Britain comparability over time is not an issue since the function of these high-stakes tests is so overwhelmingly one of ranking and selection. The main preoccupation is not with ensuring that tests and items are equally difficult in some absolute sense, but on the 'objective' rank ordering of individuals.

In contrast to the US, as Wolf points out, 'The Chinese commitment is less to the idea of standards as a measuring tool than to standards as an example and ideal'. No claims are made about substantive achievement levels or their comparability over time. Each examination is a grading system for the candidates, for example in connection with university selection, and the only relevant issue is whether the examination treats a given year's entry fairly. There appears to be an assumption that fairness is achieved by the fact that everyone is confronted by the same assessment. France is rather like China, and has centralised national examinations, the baccalaureat for example, which are very important for certification and selection purposes.

In Sweden, on the other hand, it is teachers with whom a final judgement about candidates is lodged. Consistency is an issue, but it is assumed that teachers can make such judgements so that comparability between schools and across time is maintained. There is, however, some moderation which takes place in grades 8 and 9 whereby national tests results are used to guide teachers in their own judgements. Germany has many elements of the Swedish system with teachers having a final judgement via internal marking of examinations and in many cases responsibility for setting the examinations according to State criteria.

The UK is thus rather unusual internationally in its historical concern with maintaining standards over time as well as across different curriculum subjects, and in the primacy of criterion-related concerns over norm-referencing practices. Public policy pronouncements have recently even incorporated commitments to specific test targets for future years.

Finally Wolf points out that ultimately there is no real escape from having to rely upon professional judgements in attempts to describe and

maintain 'standards', and this involves some degree of trust in those professionals. This is a persistent theme throughout the contributions to this volume. In the discussion of Wolf's paper this is emphasised by Reynolds, who also suggests that increased competition between schools may lead to the undermining of such trust. Whitburn notes how different systems vary in the extent to which they carry out centralised testing of all children, Britain being especially notable for the large amount that is carried out. She asks what the purpose of all this testing is, and whether the purposes might not be better achieved by other means. She points out that some tests are now being used primarily to make comparisons between schools and teachers and asks why it is that in England our obsession is with comparing *school* performance? 'Does it reflect our *mistrust* of our schools and our teachers which has been fuelled by those in influential positions? Or is it more a reflection of our unwillingness to attribute individual responsibility for achievement (or failure)? In Japan, there is a widespread belief in the importance of effort rather than innate ability and *pupils* are encouraged to believe that 'If you work hard enough and persevere, you can succeed.' In England the message is that *teachers* need to work harder and persevere in order for their pupils to succeed and where pupils do not achieve well, it is poor teaching that is held to be responsible.'

Aldrich begins his historical review by considering the various definitions and understandings that have been attached to the term 'standard' and tracing the usage through to the present day where it has become a touchstone of Government education policy and a term that is used, often loosely, in a great deal of public debate. He emphasizes that there is a crucial distinction between the notion of a standard as a yardstick for judging performance and a standard in the sense of the average level of attainment as measured by that yardstick. Public pronouncements often confuse these two senses.

Aldrich traces broad historical changes in levels of attainment with respect to literacy for which some kinds of generally accepted norms or yardsticks are available. These allow large changes, such as those involving the numbers of people engaged in reading, to be roughly measured and understood. Like other contributors to the symposium he also stresses the fact that there are severe definitional problems and that agreement about small or subtle changes in attainment are very difficult, if not impossible, to determine. This is evident during the second half of the twentieth century where there is much debate but little

general agreement about changes in levels of literacy, and importantly, about possible reasons for any such changes.

In the second section of his paper Aldrich discusses the latter part of the 19th century in England when the 'payments by results' system was in place. He charts the introduction of a rigid student assessment system and how opposition to it grew. Many of the debates at that time prefigure contemporary debates. These debates included issues about comparing schools with very different pupil intakes, about how *minimum* achievement targets turned into *optimal* targets for achievement, about how the most and least able were neglected in pursuit of high 'pass rates', and how creativity was discouraged. In addition there was concern that the system was conducive to a 'commercialisation' of education which was harmful. Eventually the system collapsed, although some of its assumptions about 'standards' persisted. In his conclusions Aldrich suggests that the imposition of 'quick fixes' to change 'standards' is not the way truly to raise standards and that the evidence from history supports this view.

In her discussion of Aldrich, Sutherland distinguishes between a high standard which only a few will reach and a minimum standard, and traces how these separate uses of the term developed historically for different purposes. In the introduction of examinations into the Universities and the Civil Service, standards were viewed as a fixed reference point associated with high achievement. By contrast, in the implementation of the 1862 revised code, standards were seen as defining minimum achievements. Sutherland notes how opponents of the revised code, notably Matthew Arnold, associated the imposition of crude standards with a market consumer model of education. Sutherland suggests that analysis of previous debates can often raise useful questions to ask about contemporary issues.

Prais, like Sutherland, emphasises the importance of whether a standard is meant to cater for high or low achievers and discusses how any choice is related to teaching and learning. He also makes the point that a concentration on raising *average* achievements often tends to ignore associated changes in the *spread* of achievement. He suggests that curriculum and teaching changes may have a differential effect on low and high achieving pupils.

The final discussant of Aldrich's paper, Heath, looks at evidence from the General Household Survey in order to study changes in formal qualifications during the 20th century. He shows that improvements as measured by public examination results first occurred at the lowest

levels of attainment and this reinforces the point made by Aldrich and Sutherland that even with the end of 'payments by results' the 19th century concern with achieving minimum standards persisted into the 20th century. In using changes in public examination grade distributions Heath acknowledges Cresswell's point that such grades do not represent absolute fixed standards and that any inferences have to rely on judgement. He goes further and argues that we should not expect a certification examination, which over time caters for different groups, to maintain the same underlying standards. He argues, nevertheless, that using the available evidence, real changes in attainment have taken place.

In the third section Cresswell argues strongly that examination standards cannot have the same level of objectivity and hence comparability as measurements in other sciences. They rely upon judgements of examiners and, while great care is taken in making those judgements, they are ultimately subjective. Examination 'standards' are accepted because examiners are trusted to make such judgements. Cresswell discusses the ways in which examiners go about their tasks and shows how all of their procedures, including the statistically based ones, ultimately rely upon subjective, albeit informed, judgement.

He argues that we should cease attempting to use examination results as a way of monitoring standards, but does suggest that a study of the way such things as examination formats and marking schemes have changed over time can provide interesting insights into how general perceptions of 'standards' may have changed.

In his discussion of Cresswell's paper, Gray suggests that a study of examiners themselves would be of interest. How are they selected; how do they maintain their professional status and how do they go about securing consensus? He suggests that examiners may need to take on board more external evidence in their quest for comparability. Such evidence may involve observations about changing student compositions, and also curriculum and assessment policies which may be politically influenced. Paterson emphasises the social construction and use of assessment judgements. He illustrates this with reference to social norms concerned with 'impartiality' and applies it to criterion referencing procedures. He characterises an exam system as a social institution continually seeking ways to allocate candidates to social roles, and illustrates his views by reference to differences between the Scottish and English exam systems. He points strongly to a need to carry out more research into the social relevance of

examinations. Halsey takes a broad view of the role of examinations in modern society pointing out that some form of examination seems to be required wherever a level of competence is needed for a job. He raises questions about focussing on examinations as meritocratic selection devices.

In the final section, Bartholomew's paper explores the requirements for satisfactory measurement, starting from the proposition that there must be a fundamental requirement that agreement is reached about the way in which such standards are to be measured. He points out that there is no natural unit of measurement available and one has to be constructed. His starting point is that the quantities people are interested in, such as reading achievement, are not directly observable, and that the standard approach is therefore to use things which are observable, such as responses to specific questions, as *indicators* of the underlying attribute. The measurement process then consists in combining these indicators in suitable ways to provide an *estimate* of the quantity of interest.

Bartholomew points out that the choice of indicators is important and potentially contentious, but his concern is rather with how the responses to such indicators are combined into a measurement scale. He approaches this by envisaging a statistical *model* whose role is to relate the observed responses to the assumed underlying attribute(s) and hence to use the responses to provide estimates, for individuals and groups, of that attribute. He points out the advantages of such an approach, in that it allows various assumptions to be tested and provides a set of tools for further exploring relationships between the attributes of interest and other variables. Most important of all, it allows individuals to be distinguished by their positions along a scale, reflecting the assumption that there are indeed real differences among individuals in the attribute of interest. Any statistical model also allows us the possibility of estimating the precision with which individual or group scale values can be determined—the 'reliability' of the measuring instrument.

The broad class of models Bartholomew discusses are known as 'latent trait' or 'item response' models and he discusses how these can be formulated, how to explore their dimensionality (the number of underlying attributes) and the limitations associated with this kind of modelling. He explains very clearly how any particular statistical model can be judged by comparing its predictions against the responses actually obtained from a large random sample of respondents taking

a particular test. He shows, however, that things are not always simple; often data do not allow us to distinguish between two very different models and a wide variety of assumptions may all be perfectly compatible with what is observed. He points out, however, that even though alternative explanations are possible, each may provide useful insights into individual attributes and how they interrelate. In particular he argues that some of the criticisms of mental testing have failed to understand this issue.

Finally Bartholomew considers whether a modelling perspective has something useful to contribute to debates about changing standards and presents a simple model to illustrate the real difficulties associated with making definitive statements about changes over time because we cannot separate out all the factors which are involved. He argues that the advantage of a modelling approach is that it makes clear just where the difficulties arise and hence why we can or cannot make the inferences we wish.

In his discussion of Bartholomew's paper, Goldstein looks at different possible ways of conceptualising standards and what a particular kind of definition implies for the possibility of studying differences across populations and across time. He describes two possible types; a 'constructionist' and a 'Platonic' standard. A constructionist standard is simply defined by the score on a well-specified measuring instrument. Such a score may be derived, for example, from a statistical model such as described by Bartholomew or by simply counting correct responses. What is required is agreement about how to construct and assess questions or items and how to sample individuals, and Goldstein points out some of the problems associated with such a procedure, and suggests that it is generally unattractive.

The Platonic standard is associated with attempts to conceptualise an underlying, but unobservable, attribute, which is *approximated* by a real measuring instrument; Bartholomew's discussion of constructing indicators relevant to such an attribute would be one way of operationalising this. Goldstein emphasises that what is always required is a *judgement* about how well any real instrument does in fact approximate the attribute and points out that there will generally be no agreement on this, even though some consensus may have to be reached, as in the case of public examinations. Thus, considerations of whether tests become dated over time, or whether an exam in one year measures essentially the same attributes as one in a previous year, are essentially matters for human judgement and disagreement. This leads on to a discussion of

the basic weakness of Platonic standards, namely that there is no objective way of knowing whether, over time or across populations, the approximations involved are comparable or very different. He therefore echoes many of the conclusions reached by Cresswell on the subjective nature of attempts to maintain standards.

The other discussant, Plewis, reviews some of the purposes to which educational test scores can be put. He reinforces the point made by Bartholomew about the nature of the assumptions that have to be made when making comparisons over time and makes a case for studying 'second-order' changes; he argues that a study of how *inequalities* change over time may be the key matter of concern. He makes a plea for more research into the characteristics of the current National Curriculum tests on the grounds that the issues discussed in the symposium should be much better understood by those responsible for introducing and using this assessment system.

Two abiding themes seem to emerge from this set of contributions to the debate on standards in education. The first is that the very notion of a 'standard' has to be viewed in its historical and social context. Different countries have widely varying views of what constitutes a 'standard' and how necessary such a concept is for the adequate functioning of its educational system. The theme of 'trust' between educators and the public is a recurring topic here.

The second theme to emerge strongly is that it is difficult, if not impossible, to arrive at an 'objective' definition of educational standards. Despite claims to the contrary, ultimately the final appeal is to human judgement and no amount of technical sophistication can alter this. The notion of absolute standards may be attractive for many purposes, and it may also be necessary often to act *as if* comparability over time and space really did exist. Nevertheless, it is also important to recognise the inherent limitations associated with attempts to ascribe standards. Policies based upon comparisons of examinations, tests or other devices should therefore be seen for what they really are, human judgements which, however conscientiously pursued, are ultimately subjective and influenced by culture, personality and general perceptions of the external world.

Bibliography

Adams, J. (1912). *The Evolution of Educational Theory* (London, Macmillan).

AIE (1996). *Assessment in Education*, **3**(2).

Aldrich, R. (1995). *School and Society in Victorian Britain: Joseph Payne and the new world of education* (New York, Garland).

Aldrich, R. (1996). *Education for the Nation* (London, Cassell).

Aldrich, R. (1997). *The End of History and the Beginning of Education* (London, Institute of Education).

Aldrich, V. C. (1963). *Philosophy of Art* (Englewood Cliffs, Prentice-Hall).

Anderson, R. D. (1995). *Education and the Scottish People* (Oxford, Oxford University Press).

Arnold, Matthew (1863). *A French Eton*, reprinted in *The Complete Prose Works of Matthew Arnold vol. ii, Democratic Education* ed. R. H. Super (Ann Arbor, University of Michigan Press, 1962), pp 262–325.

Arnott, M. (1993). Thatcherism in Scotland: an Exploration of Educational Policy in the Secondary Sector (PhD Thesis, Strathclyde University).

Ayer, A. J. (1946). *Language Truth and Logic* Second edition (London; Penguin).

Baird, J. (1998). What's in a Name? Experiments with blind marking in A-level Examinations. *Educational Research*, **40**(2), 191–202.

Baird, J. and Jones, B. (1998). Statistical analyses of examination standards: better measures of the unquantifiable? (Associated Examining Board Research Report—RAC/780).

Bardell, G.; Fearnley, A. and Fowles, D. (1984). *The contribution of graded objectives schemes in Mathematics and French* (Manchester, Joint Matriculation Board).

Barnes, B. (1974). *Scientific Knowledge and Sociological Theory* (London, Routledge and Kegan Paul).

Bartholomew, D. J. and Knott, M. (1999). *Latent Variable Models and Factor Analysis* (2nd edition) (London, Arnold).

Bartholomew, D. J. (1996). *The Statistical Approach to Social Measurement* (San Diego, Academic Press).

Beardsley, M. C. (1981). *Aesthetics: Problems in the Philosophy of Criticism* (Indianapolis, Hackett).

Beaton, A. E. and Zwik, R. (1990). *Disentangling the NAEP 1985–86 reading anomaly.* (Princeton, Educational Testing Service).

Benn, C. and Chitty, C. (1996). *Thirty Years On* (London, David Fulton).

Berger, P. and Luckmann, T. (1966). *The Social Construction of Reality* (London, Penguin).

Berry, C. (1997). *Social Theory of the Scottish Enlightenment* (Edinburgh, Edinburgh University Press).

Best, D. (1985). *Feeling and Reason in the Arts* (London, Allen & Unwin).

Bierhoff, H. (1996). Laying the foundation of numeracy: a comparison of primary

school textbooks in Britain, Germany and Switzerland. *Teaching Mathematics and Its Applications*, **15**, 141–60.

Billington, R. (1988). *Living Philosophy: An Introduction to Moral Thought* (London, Routledge).

Bourdieu, P. (1989). *La Noblesse d'État: Grandes Écoles et Esprit de Corps* (Paris, Les Editions de Minuit).

Brock, M.G. and Curthoys, M.C. (1998). (eds.). *The History of the University of Oxford vol. vi, Nineteenth-Century Oxford, Part 1* (Oxford, Clarendon Press).

Brooks, G. (1997). Trends in standards of literacy in the United Kingdom, 1948–1996 (paper presented at the UK Reading Association conference, University of Manchester, July 1997, and at the British Educational Research Association conference, University of York, September 1997).

Brown, A., McCrone, D., Paterson, L. and Surridge, P. (1998). *The Scottish Electorate* (London, Macmillan).

Burnhill, P., Garner, C. and McPherson, A. (1990). Parental education, social class and entry to higher education, 1976–1986. *Journal of the Royal Statistical Society*, series A, **153**, 233–248.

Burstein, J., Kaplan, R., Wolff, S., and Chi, L. (1997). Using Lexical Semantic Techniques to Classify Free-Responses (Princeton N.J. Educational Testing Service Research Report available on ETSnet at http: //www.ets.org/research/siglex.html).

Christie, T. and Forrest, G. M. (1981). *Defining Public Examination Standards* (London, Schools Council/Macmillan).

Cipolla, C. M. (1969). *Literacy and Development in the West* (London, Penguin).

Clanchy, M. (1979). *From Memory to Written Record: England 1066–1307* (London, Edward Arnold).

Collins, R. (1979). *The Credential Society* (New York, Academic Press).

Committee of Council on Education (1863). *Report of the Committee of Council on Education 1862–63* (London).

Committee of Council on Education (1872). *Report of the Committee of Council on Education 1871–72* (London).

Committee of Council on Education (1873). *Report of the Committee of Council on Education 1872–73* (London).

Committee of Council on Education (1883). *Report of the Committee of Council on Education 1882–83* (London).

Copeland, J. (1993). *Artificial Intelligence: A Philosophical Introduction* (Oxford, Blackwell).

Cox, C. B. and Dyson, A. E. (1971). (eds.). *The Black Papers on Education* (London, Davis-Poynter).

Cresswell, M. J. (1987). Describing Examination Performance: grade criteria in public examinations. *Educational Studies*, **13**(3), 247–65.

Cresswell, M. J. (1990). Gender Effects in GCSE—Some Initial Analyses (Paper prepared for a Nuffield Seminar at University of London Institute of Education on 29 June 1990) (Unpublished Associated Examining Board Research Report—RAC/517).

Cresswell, M. J. (1994). Aggregation and Awarding methods for National Curriculum

Assessments in England and Wales: a comparison of approaches proposed for Key Stages 3 and 4. *Assessment in Education*, **1**(1), 45–61.

Cresswell, M. J. (1995). Technical and Educational Implications of using Public Examinations for Selection to Higher Education. In T. Kellaghan (ed.), *Admission to Higher Education: Issues and Practice* (Dublin, Educational Research Centre and Princeton, International Association for Educational Assessment).

Cresswell, M. J. (1996). Defining, Setting and Maintaining Standards in Curriculum Embedded Examinations: Judgemental and Statistical Approaches. In H. Goldstein and T. Lewis (eds.). *Assessment: Problems, Developments and Statistical Issues* (London, Wiley).

Cresswell, M. J. (1997a). *Examining Judgements: Theory and Practice of Awarding Public Examination Grades* (PhD thesis, University of London Institute of Education).

Cresswell, M. J. (1997b). Can Examination Grade Awarding be Objective and Fair at the Same Time? Another Shot at the Notion of Objective Standards (Unpublished Associated Examining Board Research Report—RAC/733).

Cresswell, M. J. and Houston, J. G. (1991). Assessment of the National Curriculum—some fundamental considerations. *Educational Review.* **43**, 63–78.

Cressy, D. (1980). *Literacy and the Social Order: reading and writing in Tudor and Stuart England* (Cambridge, Cambridge University Press).

Damasio, A. R. (1995). *Descartes Error: Emotion, Reason and the Human Brain* (London, Papermac).

Davis, E. (1993). *Schools and the State* (London, Social Market Foundation).

Dean, C. (1998). Standards are not parents' top priority. *Times Educational Supplement*, 9 October.

Dearing, R. (1995). *Review of the 16–19 qualifications* (London, Department of Education).

Dennett, D. (1993). *Consciousness Explained* (London, Penguin).

Department for Education and Employment (DfEE). (1997). *Excellence in Schools* (London, Stationery Office).

Department of Education and Science (1967). *Children and Their Primary Schools. A Report of the Central Advisory Council for Education (England).* ii (London, DES).

Devine, M., Hall, J., Mapp, J. and Musselbrook, K. (1996). *Maintaining Standards: Performance at Higher Grade in Biology, English, Geography and Mathematics* (Edinburgh, Scottish Council for Research in Education).

Devlin, K. (1997). *Goodbye Descartes: The End of Logic and the Search for a New Cosmology of the Mind* (New York, Wiley).

Dore, R. (1996). *The Diploma Disease.* 2nd edition (London, Institute of Education).

Dreyfus, H. L. (1992). *What computers still can't do: a critique of artificial reason.* (Cambridge Mass., MIT Press).

Eagleton, T. (1993). *Literary Theory: An Introduction* (Oxford, Blackwell).

Eiser, J. R. (1990). *Social Judgement* (Milton Keynes, Open University Press).

Elwood, J. and Comber, C. (1996). *Gender differences in examinations at 18+* (London, Institute of Education).

Firestone, W. A. (1998). A Tale of Two Tests: Tensions in Assessment Policy. *Assessment in Education*, **5**(2), 175–192.

Fletcher, S. (1980). *Feminists and Bureaucrats. A study in the development of girls' education in the nineteenth century* (Cambridge, Cambridge University Press).

Fogelin, R. J. (1967). *Evidence and Meaning: Studies in Analytic Philosophy* (London, Routledge).

Forrest, G. M. and Orr, L (1984). *Grade Characteristics in English and Physics* (Manchester;,Joint Matriculation Board).

Foxman, D., Ruddock, G. and McCallum, I. (1990). *APU mathematics monitoring 1984–88 (Phase 2)* (London, Schools Examination and Assessment Council).

Fremer, J. (1989). Testing Companies, Trends and Policy Issues: A current view from the testing industry. In B. R. Gifford (ed.), *Test Policy and the Politics of Opportunity Allocation: The Workplace and the Law* (Boston, Kluwer).

French, S., Slater, J. B., Vassiloglou, M. and Willmott, A. S. (1987). *Descriptive and Normative Techniques in Examination Assessment* (Oxford, UODLE).

Galton, M. (1998). Back to consulting the ORACLE. *Times Educational Supplement*, 3 July.

Gierl, M. J. and Rogers, W. J. (1996). Factor analysis of the Test Anxiety Inventory using Canadian high school students. *Educational and Psychological Measurement*, **56**, 315–324.

Goldstein, H. (1983). Measuring Changes in Educational Attainment Over Time: Problems and Possibilities. *Journal of Educational Measurement*, **20**, 369–78.

Goldstein, H. (1995). *Interpreting International Comparisons of Student Achievement* (Paris, UNESCO).

Goldstein, H. (1996a) (ed.). *Assessment in Education, 3,* 2. Special Issue: The IEA Studies.

Goldstein, H. (1996b). International Comparisons of Student Achievement. In Little and Wolf (1996).

Goldstein, H. (1999). Performance Indicators in Education. In D. Dorling and S. Simpson (eds.). *Statistics in Society* (London, Arnold).

Goldstein, H. and Cresswell, M. J. (1996). The comparability of different subjects in public examinations: a theoretical and practical critique. *Oxford Review of Education*, **22**(4), 435–42.

Goldstein, H. and Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology*, **42**, 139–167.

Good, F. J. and Cresswell, M. J. (1988a). *Grading the GCSE* (London, Secondary Examinations Council).

Good, F. J. and Cresswell, M. J. (1988b). *Differentiated Assessment: Grading and Related Issues* (London, Secondary Examinations Council).

Gould, S.J. (1984). *The Mismeasure of Man* (London, Penguin).

Gray, J., Hopkins, D., Reynolds, D., Wilcox, B., Farrell, S. and Jesson, D. (1999). *Improving Schools: Performance and Potential* (Milton Keynes, Open University Press).

Gray, J., McPherson, A. and Raffe, D. (1983). *Reconstructions of Secondary Education* (London, Routledge).

Green, A., Leney, T. and Wolf, A. (1997). *Convergences and Divergences in European Education and Training Systems* (Brussels, EC Directorate-General XXII (Education, Training and Youth)).

Green, A., Wolf, A. and Leney, T. (1999). *Convergence and Divergence in European Education and Training Systems* (London, Institute of Education).

Hacking, I. (1965). *The Logic of Statistical Inference* (Cambridge, Cambridge University Press).

Hacking, I. (1990). *The Taming of Chance* (Cambridge, Cambridge University Press).

Hambleton, R. K. and Zaal, J. N. (eds.) (1991). *Advances in Educational and Psychological Testing* (Boston, Kluwer).

Hargreaves, D. H. (1996). Teaching as a research-based profession: policies and prospects (Teacher Training Agency annual lecture).

Heath, A. F. and Clifford, P. (1990). Class inequalities in education in the twentieth century. *Journal of the Royal Statistical Society*, series A, **153**, 1–16. .

Holland, P. W. and Rubin, D. B. (1982). *Test Equating* (New York, Academic Press).

Hollis, M. and Lukes, S. (1982). (eds). *Rationality and Relativism* (Oxford, Blackwell).

Holmes, E. (1911). *What Is and What Might Be* (London, Constable).

Jencks, C. (1972). *Inequality: A Reassessment of the Effect of Family and Schooling in America* (New York, Basic Books).

Johnson, V. E. (1997). An alternative to the traditional GPA for evaluating student performances. *Statistical Science*, **12**, 251–278.

Kelly, A. (1976). A study of the comparability of external examinations in different subjects. *Research in Education*, **16**, 37–63.

Kilpatrick, J. and Johansson, B. (1994). Standardised Mathematics Testing in Sweden: The Legacy of Frits Wigforss. *Nordic Studies in Mathematics Education*, **1**, *6–30.*

Koretz, D., Broadfoot, P. and Wolf, A. (1998) (eds.). *Assessment in Education*, **5**(3) (Special Issue on Portfolios and Records of Achievement).

Kuhn, T. (1970). *The Structure of Scientific Revolutions*, second edition (Chicago, University of Chicago Press).

Lakatos, I. (1974). *Proofs and Refutations: the Logic of Mathematical Discovery* (Cambridge, Cambridge University Press).

Little, A. (1996) (ed.). *Assessment in Education*, *4*(1) (Special Issue: The Diploma Disease Twenty Years On).

Little, A., Wang Gang, and Wolf, A. (1995) (eds.). *Sino-British Perspectives on Educational Assessment* (London, ICRA, Institute of Education).

Little, A. and Wolf, A. (1996) (eds.). *Assessment in Transition: Learning, monitoring and selection in international perspective* (Oxford, Pergamon).

Long, H. A. (1985). Experience of the Scottish Examinations Board in developing a grade-related criteria system of awards (Paper presented at the 11th annual conference of the International Association for Educational Assessment held in Oxford, England).

Macaulay, Lord (1898). *Collected Works*, 12 vols. (London, Longmans Green).

Mackenzie, D. A. (1981). *Statistics in Britain 1865–1930. The Social Construction of Scientific Knowledge* (Edinburgh, Edinburgh University Press).

McKenzie, D. (1994). The irony of educational review. *New Zealand Annual Review of Education*, **4**, 247–59.

McLean, L. D. (1996). Large-Scale Assessment Programmes in Different Countries

and International Comparisons. In H. Goldstein and T. Lewis (eds.), *Assessment: Problems, Developments and Statistical Issues* (Chichester, Wiley).

McPherson, A. and Willms, J. D. (1987). Equalisation and improvement: some effects of comprehensive reorganisation in Scotland. *Sociology*, **21**, 509–39.

Madaus, G. and Raczek, A. (1996). Turning Point for Assessment: Reform Movements in the United States. In Little and Wolf (1996).

Menet, J. (1874). *A Letter to a Friend on the Standards of the New Code of the Education Department* (London, Rivingtons).

Morrison, H. G., Busch, J. C. and D'arcy, J. (1994). Setting reliable national curriculum standards: a guide to the Angoff procedure. *Assessment in Education*, **1**, 181–199.

Murphy, R. J. L. (1982). Sex differences in Objective Test performance. *British Journal of Educational Psychology*, **52**, 213–19.

Murphy, R. J. L., Burke, P., Cotton, T., Hancock, J., Partington, J., Robinson, C., Tolley, H., Wilmut, J. and Gower, R. (1996). *The Dynamics of GCSE Awarding: Report of a project conducted for the School Curriculum and Assessment Authority* (London, SCAA).

Newcastle Report (1861). *Report of the Commissioners appointed to inquire into the State of Popular Education in England*, PP 1861 XXI (ii) (London).

Newton, P. (1996). The reliability of marking of GCSE scripts: Mathematics and English. *British Educational Research Journal*, **22**, 405–20.

Newton, P. (1997a). Measuring comparability of standards between subjects: why our statistical techniques do not make the grade. *British Educational Research Journal*, **23**(4), 433–49.

Newton, P. (1997b). Examining Standards Over Time *Research Papers in Education*, **12**(3), 227–48.

Orr, L. and Forrest, G. M. (1984). *Investigation into the relationship between grades and assessment objectives in History and English examinations* (Manchester, Joint Matriculation Board).

Orr, L. and Nuttall, D. L. (1983). *Determining Standards in the Proposed Single System of Examinations at 16+* (London, Schools Council).

Paterson, L. (1992). The influence of opportunity on aspirations among prospective university entrants from Scottish schools, 1970–1988. *Statistics in Society, Journal of the Royal Statistical Society*, series A, **155**, 37–60.

Paterson, L. (1995). Social origins of under-achievement among school-leavers. In L. Dawtrey, J. Holland, M. Hammer and S. Sheldon (eds.), *Equality and Inequality in Education Policy* (Milton Keynes, Open University Press).

Paterson, L. (1997). Student achievement and educational change in Scotland, 1980–1995. *Scottish Educational Review*, **29**, 10–19.

Paterson, L. (1998). The Scottish parliament and Scottish civil society: which side will education be on? *Political Quarterly*, **69**, 224–33.

Paterson, L. (forthcoming). Scottish traditions in education. In H. Holmes (ed.), *Compendium of Scottish Ethnology, vol. 11* (Edinburgh, Scottish Ethnological Research Centre).

Paterson, L. and Raffe, D. (1995). Staying on in full-time education in Scotland. *Oxford Review of Education*, **21**, 3–23.

Payne, J. (1872). 'Why are the Results of our Primary Instruction so Unsatisfac-

tory?', *Transactions of the National Association for the Promotion of Social Science.*

Phillips, M. (1996). *All Must Have Prizes* (London, Little, Brown and Company).

Pirsig, R. M. (1974). *Zen and the Art of Motorcycle Maintenance: An Inquiry into Values* (London, Bodley Head).

Plewis, I. (1998). Inequalities, Targets and Zones. *New Economy*, **5**, 104–8.

Plewis, I. (1999). What's Worth Comparing in Education? In D. Dorling and S. Simpson (eds.). *Statistics in Society* (London, Arnold), 273–80.

Pole, D. (1961). *Conditions of Rational Inquiry: A Study in the Philosophy of Value* (London, Athlone).

Power, M. (1997). *The Audit Society: Rituals of Verification* (Oxford, Oxford University Press).

QCA (1998). *GCSE and GCE A/AS code of practice* (London, Qualifications and Curriculum Authority).

Reynolds, D., Creemers, B. P. M., Stringfield, S. and Teddlie, C. (1998). Climbing an educational mountain: conducting the International School Effectiveness Research Project. In G. Walford, *Doing research about education* (Lewes, Falmer Press).

Roach, J. P. C. (1971). *Public Examinations in England 1850–1900* (Cambridge, Cambridge University Press).

Robertson, C. (1992). Routes to higher education in Scotland. *Scottish Educational Review*, **24**: 3–16.

Rose, S. (1997). *Lifelines, Biology, Freedom, Determinism* (London, Penguin).

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education*, **13**, 191–209.

Schools Council (1979). *Standards in Public Examinations: Problems and Possibilities*, Report from the Schools Council Forum on Comparability (London, Schools Council).

SEC (1984). *The development of Grade-related Criteria for the General Certificate of Secondary Education—a briefing paper for working parties* (London, Secondary Examinations Council).

SEC (1985). *Reports of the Grade-related Criteria Working Parties* (London, Secondary Examinations Council).

SEC (1986). Draft Grade Criteria. *SEC News Number 2* (London, Secondary Examinations Council).

SEC (1987). Grade Criteria—Progress Report. *SEC News Number 6* (London, Secondary Examinations Council).

Shavit, Y. and Blossfeld, H. P. (1993). *Persistent Inequality: Changing Educational Attainment in Thirteen Countries* (Boulder, Col., Westview Press).

Skolöverstyrelsen (1980). Quoted in J. Kilpatrick and B. Johansson (1994). Standardised Mathematics Testing in Sweden: The legacy of Frits Wigforss. *Nordic Studies in Mathematics Education*, **1**, 6–30. .

Smith, J. V. and Hamilton, D. (1980) (eds). *The Meritocratic Intellect* (Aberdeen, Aberdeen University Press).

Start, B. and Wells, K. (1972). *The trend of reading standards* (Slough, National Foundation for Educational Research).

Stedman, L. C. (1998). An Assessment of the Contemporary Debate over US

Achievement. In D. Ravitch (ed.), *Brookings Papers on Education Policy* (Washington DC, Brookings Institution Press), 53–119.

Stephens, W. B. (1987). *Education, Literacy and Society, 1830–70: the geography of diversity in provincial England* (Manchester, Manchester University Press).

Sutherland, G. (1973a). *Policy-Making in Elementary Education 1870–1895* (Oxford, Clarendon Press).

Sutherland, G. (1973b) (ed.). *Matthew Arnold on Education* (London, Penguin).

Sutherland, G. (1984). *Ability, Merit and Measurement. Mental testing and English education 1880–1940* (Oxford, Clarendon Press).

The Scotsman Education (1998). 30 September: 4–5.

Thom, D. (1986). The 1944 Education Act: the 'art of the possible. In Harold L. Smith (ed.), *War and Social Change: British Society in the Second World War* (Manchester, Manchester University Press), 101–28.

Vincent, D. (1989). *Literacy and Popular Culture: England 1750–1914* (Cambridge, Cambridge University Press).

Walden, G. (1996). *We Should Know Better: solving the educational crisis* (London, Fourth Estate).

Wang Binhua (1995). Comparing HSCE in the People's Republic of China and GCSE in England. In Little, Wolf and Wang Gang (1995).

Wang Gang (1995). The Development of Public Educational Examinations in China from 1980. in Little, Wolf and Wang Gang (1995).

Wiliam, D. (1996a). Meanings and Consequences in Standard Setting. *Assessment in Education*, **3**(3), 287–307.

Wiliam, D. (1996b). Standards in examinations: a matter of trust? *The Curriculum Journal*, **7**(3), 293–306.

Wilmut, J. and Rose, J. (1989). *The Modular TVEI Scheme in Somerset: its concept, delivery and administration* (Report to the Training Agency of the Department of Employment, London).

Wolf, A. (1995). Competence Based Assessment (Buckingham, Open University Press).

Wolf, A. and Steedman, H. (1998). Basic Competence in Mathematics: Swedish and English 16 year olds. *Comparative Education*, **34**, 3.

Wood, R. (1991). *Assessment and Testing: A survey of research* (Cambridge, Cambridge University Press).

Young, M. (1958). *The Rise of the Meritocracy 1870–2033* (London, Penguin).