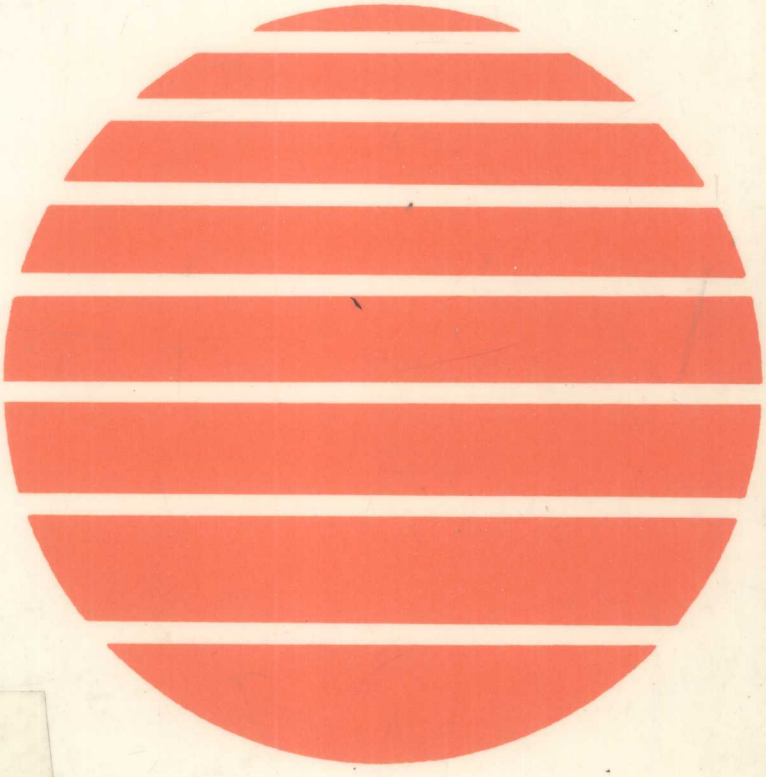
The Falmer Press
Contemporary Analysis in Education

ASSESSING EDUCATIONAL ACHIEVEMENT



Edited by Desmond L. Nuttall

Contemporary Analysis in Education Series

Assessing Educational Achievement

Edited by
Desmond L. Nuttall

The Falmer Press
(A member of the Taylor & Francis Group)
London and Philadelphia

UK The Falmer Press, Falmer House, Barcombe, Lewes, East Sussex, BN8 5DL

USA The Falmer Press, Taylor & Francis Inc., 242 Cherry Street, Philadelphia, PA 19106-1906

First published 1986

Library of Congress Cataloging in Publication Data

Main entry under title:

Assessing educational achievement.

(Contemporary analysis in education series)
Bibliography: p.
Contents: Assessing for learning / Harry Black — The prospects for public examinations in England and Wales / Henry G. Macintosh — Australian examination systems / Graeme Withers — [etc.]
1. Educational tests and measurements — Addresses, essays, lectures. I. Nuttall, Desmond L. II. Series.
LB3051.A766 1986 371.2'6 85-20679
ISBN 1-85000-056-5 (pbk.)

Typeset in 11/13 Garamond by
Imago Publishing Ltd, Thame, Oxon

Jacket design by Leonard Williams

Printed in Great Britain by Taylor & Francis (Printers) Ltd, Basingstoke

General Editor's Preface

The assessment of educational achievement is at times more than a reasonable public concern: it verges on public hysteria. Part, though not all, of the reason for this is a failure to grasp the nature and limitations of educational assessment. Even among educators there is only a limited grasp of the complexities of assessing educational achievement, and about some forms of assessment there is simply prejudice.

This highly professional and very readable collection of essays, skilfully edited by Desmond Nuttall, should go a long way to setting matters right and placing them in perspective. Moreover, the collection shows that there is an international dimension to problems of assessing educational achievement, at least in the English speaking world.

What the collection also amply illustrates is that there are many inventive and educationally committed minds at work in the field of educational assessment. If this could be fully appreciated by those who demand the assessment of educational achievement, then the best effects of assessing achievement could be realized and the worst mitigated.

Philip Taylor
Birmingham
December 1985

Contents

General Editor's Preface vii

Editorial Introduction
Desmond L. Nuttall, Professor of Educational Psychology, The Open University 1

Section 1
Assessment for Learning
Harry Black, Research Officer, Scottish Council for Research in Education 7

The Prospects for Public Examinations in England and Wales
Henry G. Macintosh, Secretary, Southern Regional Examinations Board 19

Australian Examination Systems: Eight States of the Art
Graeme Withers, Senior Research Officer, Australian Council for Educational Research 35

Alternatives to Public Examinations
Patricia M. Broadfoot, Lecturer in Education, University of Bristol 54

Section 2
Exams in Context: Values and Power in Educational Accountability
Stewart Ranson, Institute of Local Government Studies, University of Birmingham, John Gray, David Jesson and Ben Jones, University of Sheffield 81

A Critique of the APU
Caroline Gipps, Senior Research Officer, Screening and Special Educational Provision in Schools Project 99

Testing in the USA
Archie E. Lapointe, Executive Director, National Assessment of Educational Progress 114

Educational Assessment in the Canadian Provinces
Les D. McLean, Professor and Head, Educational Evaluation Centre, The Ontario Institute for Studies in Education 125

Section 3
Problems in the Measurement of Change
Desmond L. Nuttall, Professor of Educational Psychology, The Open University 153

Models for Equating Test Scores and for Studying the Comparability of Public Examinations
Harvey Goldstein, Professor of Statistical Methods, University of London Institute of Education 168

The Agenda for Educational Measurement
Robert Wood, Senior Research Fellow, Flinders University 185

The Contributors 205

Editorial Introduction

About three years ago, an issue of *Educational Analysis* (Volume 4 Number 3) edited by me was published under the title 'Assessing Educational Achievement'. I was very pleased when the publishers invited me to edit a second collection, since there have been many important developments in the intervening period. I was equally pleased when all the original contributors (save one) agreed to update their contributions in the light of subsequent developments in their fields. Many have radically revised their chapters to reflect the rapid pace of change in national policies, and all have brought them up to date by incorporating new information, new data and new references. The only original contributor who did not feel able to update his article was Roy Forbes, then of the Education Commission of the States, the body responsible for the National Assessment of Educational Progress (NAEP). In 1983, the Education Commission of the States was unsuccessful in its tender to take on the next phase of NAEP, and Educational Testing Service won the contract. Appropriately, the new Executive Director of NAEP, Archie E. Lapointe, kindly agreed to write about American developments.

The assessment of educational achievement serves many functions in education. In one guise, it is an integral part of teaching and learning, though most assessment is carried out informally — through questions and answers in class, through observation of students at work — rather than through the formal and artificial means of tests and examinations. Yet it is these more formal types of assessment that inevitably catch the public eye and generate debate and research (the latter occasionally informing the former). The research has usually been of the technical kind, investigating the efficiency of the tests as measuring instruments and as predictors of future success; until recently, relatively little research had been

carried out on their social, psychological and educational effects, but such as has been done aligns with the mood of the times to question many of the forms and functions of assessment in education.

The chapters in this volume contribute to this process of questioning. They fall into three main sections: the first section comprises four chapters that look at the assessment of the individual and the second, four chapters that review the use of the assessment of individuals as a way of assessing the performance of educational institutions or the educational system as a whole. The final section consists of three chapters that are more theoretical, grappling with problematic conceptual issues in the field of educational assessment.

In the first chapter, Harry Black, drawing on his experience of diagnostic and school-based assessment, indicates how assessment might be developed to be of direct value to teacher and student (formative assessment) with particular emphasis upon criterion-referenced procedures. To achieve this, though, we need to break away from the dominant model of summative assessment, as epitomized in the British examination systems. These systems show signs of moving towards a greater degree of criterion-referencing; Harry Black sees danger in this move if it re-establishes an ascendancy for summative assessment. In the next chapter, Henry Macintosh examines the developments in the examination systems of England and Wales, a difficult task since there have been so many proposals for change in recent years and final decisions on many of them are imminent. On one, the common system of examining at 16+, a decision was announced in June 1984, confounding the predictions of Henry Macintosh and many others; a new examination, the General Certificate of Secondary Education, *is* to be introduced from 1988, but on terms dictated by central government that drastically reduce the autonomy of teachers. At 17+, there is still a ferment of activity but, at 18+, the dominant position of the universities seems likely to continue to stifle long-overdue reform. In Australia, too, as Graeme Withers reveals in his chapter, the universities exert a dominating influence on the curriculum and assessment at the point of transition from school, though alternatives to examinations are gaining ground, not without controversy. In contrast to Britain, though, Australian states have almost completely abolished examinations below the level of university entrance, and allow schools to develop their own courses and assessments.

In the final chapter in this section on assessment of the individual, Tricia Broadfoot points to the bugbear of assessment selection. While selection (for different forms of further and higher education as

well as for employment) is still required of the educational system (and while examinations have other important societal functions to perform, for example in controlling the curriculum), any alternative to public examinations has to fulfil these same functions. Many of the alternatives, like pupil profiles in Britain and *orientation* in France, may therefore prove to be retrograde rather than progressive. Only through alternatives that challenge the functions themselves might we find a way forward.

Assessments of individuals are frequently aggregated to provide assessments of classes (and teachers), institutions and the educational system. The current emphasis on the accountability of public institutions, not least schools, is the theme underlying the next group of four chapters. Stewart Ranson, John Gray, David Jesson and Ben Jones analyse the concept of accountability in their chapter and show how dangerously simplistic the use of public examination results to judge the effectiveness of schools can be. Moreover, the publication of examination results required under the law serves to signal and reinforce the predominant academic emphasis of British schooling, an emphasis that other initiatives in assessment documented in earlier chapters are seeking to undermine.

A little of the pressure on examination results as measures of the efficiency of the educational system was removed by the establishment, in 1974, of the Assessment of Performance Unit (for England, Wales and, later, Northern Ireland). In her chapter, Caroline Gipps offers a critique of the work of the APU showing how, often through lack of forward planning and proper analysis of the problems, it has so far failed to meet a number of its objectives, while succeeding in others. Somewhat surprisingly, perhaps its greatest success is in the production of new and imaginative methods of assessment that might be of great value to the teacher in the classroom.

Without systems of public examinations, the USA and the individual states were early in the business of testing for assessing the efficiency of the system and the institutions. The chapter by Archie Lapointe shows how concern about educational performance is as great as ever, and how demands for testing show no signs of abating. Indeed, the experience and familiarity with the use of test results over the past twenty years has created a new level of sophistication in the expectations of professional and lay publics. NAEP, under its new management led by Archie Lapointe, is responding by making a series of changes designed to enhance its effectiveness (though in my own chapter I am critical of some of these changes). In Canada, too, assessment of the performance of educational systems flourishes and

grows more complex, as the review by Les McLean demonstrates. But we do not always know what to do with the results of these testing programmes, nor do we pay enough attention, McLean feels, to the issue of validity, a topic that he examines closely and critically. My own chapter, the first in the final section, examines another thorny issue in national and local assessment, the measurement of change in performance over time. Change has to be related to a constant (a measuring instrument or a measurement scale) but the constant appears elusive.

The final two chapters are, in my view, the most important and most enduring in the collection. Harvey Goldstein's provides a clarification of models underlying the concepts of equating and comparability, the latter one of the most researched aspects of public examinations in Britain. His main conclusion (that the difficulty of an examination and its relevance to a syllabus are inherently confounded) implies that only in very special circumstances is it possible to investigate the comparability of standards empirically. Otherwise, comparability is a myth as I have argued elsewhere.[1]

The last chapter, by Bob Wood, draws together many of the issues raised in earlier chapters, such as the dominance of the tradition of individual differences and the search for technical, rather than educational, improvement. He offers a programme for educational measurement that would surely appeal to all the contributors to this volume, and indicates some of the promising lines of development that can already be detected. In common with the other chapters, but more explicitly, his chapter conducts and orchestrates the reconstruction of educational measurement that is long overdue.

Desmond L. Nuttall
The Open University

Note

1 NUTTALL, D. (1979) 'The myth of comparability', *Journal of the National Association of Inspectors and Eductional Advisers*, autumn, pp. 16–17.

Section 1

Assessment for Learning

Harry Black
Scottish Council for Research in Education

In 1980, a survey in Scotland (SCRE, 1980) showed that while 87 per cent of secondary schools claimed to have an assessment policy for reporting purposes, only 29 per cent had a policy for other non-reporting assessment. Furthermore, of the former, 40 per cent of these schools had a *written* policy, but only 36 per cent of those with non-reporting policies had it in written form. That only one in ten schools has a policy in this vital area formally committed to paper is not evidence that assessment for reasons other than reporting does not take place. But it is symptomatic of the low priority given to assessment for reasons other than reporting at an organizational level.

Other simple comparisons substantiate the point. Consider the amount of time, energy and money spent by both individual teachers, and schools in general, on setting and marking continuous assessment tests, end of session examinations and mock 'O' levels. Reflect on the money spent by examination boards and the number of assessment specialists employed by them. Read, if you can find a sabbatical term, the literature on the technology of assessment for reporting and certification. Compare these in turn with the complete lack of support normally given to teachers in devising and applying procedures to pinpoint their students' learning problems, with the virtual absence of outside agencies to develop formative assessment instruments and procedures, and the limited literature on the topic.

There can be little doubt that summative assessment, which can be defined simply as assessment of the outcomes of education for purposes of reporting or certification, dominates the educational psyche of assessment. Formative assessment, which can be thought of as comprising forms of student evaluation carried out to monitor progress with a view, where appropriate, to altering the final out-

comes, holds a position somewhere behind the closed and private door of the classroom.

Yet it would be wrong to argue that formative assessment is an innovation. Teachers have always taken account of learning difficulties in their classrooms and reacted to them. Questions are asked at a class and individual level, mistakes are noticed as the teacher wanders about the classroom. Spelling, punctuation and substantive errors are corrected in written work, and common errors are noticed when tests are marked. Teaching and learning, except in the extreme case of programmed learning, are interactive, and informal formative assessment is a vital element in the whole process.

The obvious question then is that if formative assessment has long been part of the tradition of teaching, why should one draw comparisons with summative assessment? By implication one appears to be dissatisfied with the status quo. Are there not enough bandwagons to jump on without pushing one which might interfere with sound existing teaching practice? This paper *challenges* this common stance by making four assertions.

1 There is evidence that the nature of formative assessment has long been an issue;
2 Formative assessment as practised is not all that it might be;
3 There is evidence that carefully thought out formative models can make a substantial difference to learning;
4 Sound formative assessment needs is own technology and this is *not* the developed technology presently made available to teachers.

Formative Assessment as a Longstanding Issue

There are not many names from the educational world of the mid-nineteenth century which have retained their position amongst present day scholars. Amongst the few, Charlotte Mason, a teacher in the middle-class schools of Victorian England, and Helen Pankhurst of Dalton, Massachusetts, are known for their development of child-centred education. Their challenge was essentially to the way in which students learn. But a resultant change was to the nature of assessment. The extent of the change was such as to make Taylor (1965) note that when the Dalton plan was introduced to some schools in this country 'The convention that written work was always to be corrected resulted in masters disappearing in a blizzard of paper.'

Awareness of the issues was not confined to teaching methodology. An interesting early example of formative assessment technology can be found in a letter written by the Reverend George Fisher, Principal of Greenwich Hospital School, and quoted by E. Chadwick (1864). In it he describes a scale book which provides examples of works of differing levels of attainment and which can be used as a fixed standard against which to compare the work of individual pupils. In addition to his scale for writing he states that:

> By a similar process values are assigned for proficiency in mathematics, navigation, scripture knowledge, grammar and composition, French, general history, drawing and practical science respectively. Questions in each of these subjects are contained in the 'scale book' to serve as types not only of the difficulty but of the nature of the questions for the sake of future reference.

That there existed an awareness of the importance of formative assessment from the early stages of formal education cannot be in dispute. But two things happened which turned the focus elsewhere. First, the introduction of public examinations directed the attention of teachers to goals beyond their own classrooms. Second, the development of intelligence tests, especially in the United States, turned the attention of assessment specialists to the development of instruments which would sort pupils into a normal distribution of attainment. Formative assessment, of course, continued to take place but its development was dwarfed by the burgeoning summative tradition.

The most significant development which took place in the first sixty years of this century was the construction of standardized diagnostic tests. These were largely confined to basic skills in number and language. Thus, for example, Buswell and John (1926) produced a diagnostic test in arithmetic and provided a chart and manual for the teacher. In other cases diagnostic tests were built into textbooks. But today such tests are seldom used in schools, not least perhaps because modern conceptually oriented mathematics is less easily tested in this way than the basic number skills of traditional arithmetic.

In the 1960s, however, several significant developments took place in the United States where standardized testing had become a substantial industry. Programmed learning, which was seen at the time as a bright new solution for the child-centred tradition, became the focus of considerable attention. But the essence of programmed learning was that students should successfully complete a unit of

work or 'frame' before they moved on to the next. The function of formative tests in such a system is to ascertain whether a student has, or has not, attained the intended outcome of the frame. But the test-conscious educationists who tried to apply their norm-referenced technology to the situation found that it did not work. They did not want to know who had performed best on the frame, nor to have the students spread on a normal distribution. And so the term 'criterion-referenced test' was coined by Glaser (1963) to describe instruments which would best allot students to mastery or non-mastery states.

Coincident with these developments, formative assessment was being applied by Bloom and his colleagues (1971 and 1976) in a mastery learning context. In both these cases the technology and the role of formative assessment were clearly distinguished from that of summative assessment.

Formative Assessment in British Schools

The British assessment tradition has had different priorities over the last twenty years. The influence of summative assessment and, in particular, external certificates, has had a dominant role in teacher thinking. The advent of the CSE and Mode III syllabi may have been a watershed in curriculum thinking, but its impact on formative assessment must be in doubt. In particular, teachers have developed skills which are appropriate to the construction of summative internal examinations. We have also moved en masse towards the 'progressive' notion of continuous assessment. As a result, school assessment is dominated by staccato forms of the old end-of-session examinations. Continuous assessment in action means continual examination for reporting, and to make matters worse, many teachers are doing it rather well because of the skills they have picked up from the exam boards.

The problem is that the price we have had to pay for a better summative assessment model in schools has actually reduced the likelihood of formative assessment taking a more prominent role. Black and Dockrell (1980), for example, noted that in most cases where they saw continuous assessment taking place, feedback was in the form of a general attainment grade giving no real information about specific strengths and weaknesses. Furthermore, assessment typically took place at the end of each unit of work by which time it was too late to take remedial action. In a later study (Black and Dockrell, 1984), they show that some teachers think that continuous

assessment as practised is a real obstacle to the introduction of diagnostic assessment procedures. The reason is that the pressures of carrying out systematic continuous *and* diagnostic assessment are seen as placing intolerable demands on preparation, marking and testing time.

Of course, there are many examples of interesting practice in formative assessment to be found in British schools today. Black and Broadfoot (1982), for example, give a number of case studies of formative procedures in both primary and secondary schools. But the disturbing aspect of their description is that what is described in the case studies has to be seen as innovation not just by the writers but by the teachers themselves.

In fact, it is difficult to gain a clear impression of the extent and nature of formative assessment as practised by individual teachers either in the past or today. Not only does one have to breach the defensiveness of teachers in describing what their approach may be, but observational studies are necessary to evaluate the use which is made of the data collected. There is no doubt that some teachers carry out formative assessment both effectively and conscientiously. Many others are conscientious but ineffective. Others are neither. It is also the case that teachers from different subject backgrounds have differing traditions to live up to. Equally, in some schools, a deliberate policy of requiring teachers to perform certain assessment tasks means that formative assessment is expected to be carried out. But the facts about what actually takes place are not readily available, and even when they have been gathered, are sensitive in the extreme. Three points can, however, be made.

1 There is evidence that both parents (SCRE, in progress) and pupils (Black and Dockrell, 1984) rank formative information highly in terms of the feedback they would like from assessment.
2 Where teachers are put in the situation of developing alternative formative assessment models, they typically admit that their existing procedures were not designed with individual feedback of learning problems as a high priority (Black and Dockrell, 1984).
3 There is evidence that carefully planned programmes of formative assessment can have a wide range of positive impacts on learning and teaching.