

Gipps, C. (ed.) Bedford

working paper no 29
Inst. of Education,
London.

Can Graded Assessments, Records of Achievement, and Modular Assessment Co-exist with the GCSE?

Harvey Goldstein and Desmond L. Nuttall

____ (1984b), Speech to Assistant Masters and Mistresses Association, Bournemouth, April. DES Press Release 65/84.

Kempe, R.F. and L'Odiaga, J. (1984), 'Criterion-referenced interpretation of examination grades', *Educational Research*, Vol.26, pp.56-64.

Murphy, R.J.L. (1984), 'Mission impossible', *The Times Educational Supplement*, 30 November.

Murphy, R.J.L. and Pennycook, D.B. (1986), 'Graded assessments and the GCSE' in T. Horton (ed.), *GCSE: examining the new system*. London: Harper & Row.

Nuttall, D.L. (1984), 'Doomsday or a new dawn? The prospects for a common system of examining at 16-plus' in P.M. Broadfoot (ed.), *Selection, Certification and Control*. Basingstoke: Falmer Press.

Orr, L. and Nuttall, D.L. (1983), *Determining Standards in the Proposed Single System of Examining at 16-plus*, Comparability in Examinations Occasional Paper 2. London: Schools Council.

Schools Council (1979), *Standards in Public Examinations: problems and possibilities*, Comparability in Examinations Occasional Paper 1. London: Schools Council.

Scottish Education Department (1986), *Assessment in Standard Grade Courses: proposals for simplification* (McClelland Report). Edinburgh: SED.

Secondary Examinations Council (1984), *The Development of Grade-Related Criteria for the General Certificate of Secondary Education. A briefing paper for working parties* London: SEC.

____ (1985), *Secondary Examinations Council Annual Report 1984-85*. London: SEC.

Philosophical considerations

The other chapters in this volume describe and evaluate various aspects of the GCSE. Here we look at the relationship between the aims of the GCSE and those of other major assessment initiatives — graded assessments and records of achievement. Furthermore, since the focus of this volume is the GCSE, we do not consider the possible operation of graded assessments or records of achievement independent of GCSE or as the dominant modes of assessment, although this is not to deny that either (or both in partnership) might eventually predominate at least at some levels of education.

We shall describe each of four types of assessment (graded assessments, records of achievement, modular assessments and the GCSE) in terms of their educational justifications, followed by an analysis of the implicit and explicit educational philosophies embodied in these justifications. From there we shall draw out the similarities and conflicts between the types, and hence the issues likely to emerge in any simultaneous implementation of one or more of the types of assessments in the same educational environment alongside GCSE.

A major difficulty we have found in discussing philosophical or theoretical foundations is that there is very little serious attempt to relate any of the current initiatives to existing educational philosophies or to indicate the outline of a new one. Thus, the graded assessment movement draws much of its inspiration and rhetoric from graded tests in other areas, notably in artistic and sports fields, but with minimum critical analysis of the similarities and differences between those areas and general education in schools. Records of achievement are commended to us as providers of a wider portfolio of information, without addressing the question of the optimum amount of information that is required for different purposes. In similar vein, the new GCSE, as might be anticipated from

an essentially bureaucratically driven exercise, deals with the notion of 'standards' by abandoning a critical definition in favour of metaphor.

Part of our task, therefore, has been to infer key features of the frameworks within which advocates of the different assessments implicitly seem to operate. Yet our aim is not to engage only in philosophical argument, but to use that to prepare a practical critique. If the result appears unduly critical, then we make no apology but rather invite those who do not share our conclusions to join us in a debate we feel is long overdue. We should also make it clear that we accept that the very fact of change often produces new perspectives and hence important innovations in curriculum and teaching. Undoubtedly this is true of new assessment approaches. Our concern here, however, is to examine the specific claims of these approaches rather than their effects, which are often beneficial if sometimes unplanned or unexpected.

Graded assessments

The early discussion of graded testing or assessment often involved reference to graded examinations in music or grade certificates awarded for swimming, etc. (see, for example, Harrison, 1982). The advantages of such assessments were said to be in the short-term 'positive' motivation provided by relatively clearly defined tasks to be 'mastered', or competencies to be demonstrated. The term 'positive' refers to the reinforcement provided by a high probability of 'passing' a grade, which in turn implies that students are not entered unless they are perceived to have a high probability of passing. Thus the 'positive' nature of the assessment resides in appropriate entry decisions rather than in the assessment itself, and so depends upon the actions of teachers and others. If this is so, then at the very least we need to have an understanding of how the mechanism works, if indeed it does, in other parts of the curriculum. Furthermore, what are the procedures whereby pass rates are maintained and how do these react upon teaching and learning? Such questions lead to broader ones about the relationships between tutor and student, about whether there is an inherent link between this relationship and the form of assessment, and about whether the assessment method can be 'lifted' from one context to another. Of particular interest is the question of how 'negative' achievement or 'failure' is dealt with. We shall return to these issues, and also what we call context-related assessment, below.

The second common aspect of graded assessments is the acceptance of the idea of 'mastery'. Because of its often close association, or even

identification, with the notion of criterion-referencing we shall be discussing the two together below. We note, however, that both notions relate explicitly to a body of educational writing largely concerned with the formulation of detailed, and usually behavioural, educational goals and methods of assessing the achievements of those goals (e.g., Bloom, Hastings and Madaus, 1971). In practice, however, this literature tells us little about how to operationalize notions of mastery in relation to broad-based curricula as opposed to narrow definitions of skill; consequently, the development of graded assessments has been *de novo*, rather than part of any systematic extension or application of the existing theory of mastery learning and assessment.

The use of the results of graded assessments does not seem to have been fully thought through. Apart from their obviously redundant use in deciding whether to proceed to the next 'level' of assessment (given the policy governing entry decisions), the results of many graded assessments are to be equivalenced to a grade level of GCSE. This has been unashamedly justified in order to achieve acceptance and 'respectability' for graded tests, and there has been little examination of the consequences of such equivalencing. It is tacitly assumed that it can be done without prejudicing existing aims, but this is not guaranteed and we shall examine this further when we discuss the nature of assessment instruments when they are required to perform a selection function.

Finally, an issue which afflicts all assessment procedures (perhaps more in the United Kingdom than elsewhere in the world) is that of *comparability* — across time as the content of the assessment changes, across institutions such as schools which use different content and contexts to assess the same objectives, and across domains or subjects. The only known areas where comparability or equating can be said to work are very narrowly defined cases involving costly norm-referenced procedures, hardly applicable to the schemes being developed (see Holland and Rubin, 1982; Goldstein, 1986). Yet the descriptions of graded assessment schemes are predicated on a common understanding of what each level 'means'. What is unclear is the epistemological basis for such meaning — is the intention to 'objectify' it in terms of mathematically-defined relationships as in test equating, or to refer it to expert judgement, or perhaps simply to ignore the problem by pretending it does not exist? We shall have more to say about this later.

Records of achievement

A principal distinguishing feature of records of achievement (or profiles) is their incorporation of explicit and separate assessments of a much wider

range of achievements than other modes of assessment, especially in the affective domain. It is, however, worth remarking that there appears to be no reason why such achievements should not be incorporated into the other three forms of assessment. That this appears not to have been seriously considered reflects the provenance of graded assessments, modular assessments and the GCSE and their primary concern with the same set of issues. The main appeal of profiling is to the notion of education for the whole person, to a desire to give positive value and status to affective and other non-academic achievements. For many, the real value of this appeal is to the teacher and above all to the student himself or herself and lies in the *process* of education, and the control of teacher and student over the assessment. But few deny (albeit reluctantly in some cases) that status can be conveyed only by means of a permanent, observable record negotiated between student and teacher. It is this record which, along with the other kinds of assessment, would be 'used' . . . by whoever wants to use it. Clearly, to achieve value for non-cognitive activities, knowledge of them has to be of use to someone. Yet we find little explicit discussions of who that person (other than the student) is, but there seems to be a general presumption, as in the Oxford Certificate of Educational Achievement (OCEA), that it is a selector of sorts — an employer, perhaps, or a Youth Training Scheme (YTS) managing agent. So again we find an instrumental end-product, raising the issue we mentioned above of the way in which the use of an assessment interacts with its initial philosophy. A fundamental question is how far the ideal of honest and objective recording is compatible with the reality of knowing that such recording may have a role in determining life chances. This is hardly a novel question, but it still deserves a considered response.

Two other important issues arise with records of achievement which have counterparts in the other forms of assessment. The first is made explicit in many of the profiling systems (e.g. those of the City and Guilds of London Institute) that consist of graded descriptions of each of a wide range of achievements. A key feature of these is their out-of-context formulation, that is, their phrasing in terms that sound very much like general 'skills' or 'abilities' presumed to exist independently of any particular context. Here, a direct line of descent can be traced from a traditional and widely held belief in such concepts as 'intelligence', though this belief is one that has increasingly been questioned. Yet if such descriptions convey little real information, is it then possible to generate records of achievement in terms of context-specific descriptions?

The second issue is again that of 'positive' achievement. If negative

comment is to be eschewed (as the Department of Education and Science (1984) advocates), 'negative' achievement presumably becomes that range of possible achievements which are not mentioned in the record. There would seem to be a clear invitation to read between the lines, and this leaves us wondering what the term 'positive' achievement really signifies.

Modular assessment

A number of different arguments are put forward for breaking up a course of study into relatively small units or modules, and assessing achievement at the end of each. One argument is for curricular flexibility (and sometimes organizational flexibility) so that students can have more choice, and a greater variety of potentially valuable courses can be created. But where the modules are being combined to create courses for the GCSE examination (as is the case in many TVEI schemes), the national criteria tend to create a straitjacket and much of the curricular flexibility is lost because only a very limited number of combinations of modules will be acceptable as a GCSE course in physics, say.

Some modular schemes (such as those operated by the Business and Technician Council and the Open University) consist of modules that are studied over a period as long as a year. Other schemes (e.g. Inner London Education Authority, 1984) advocate modules that can be completed in as little as six or eight weeks, so that students receive relatively rapid public feedback about performance. In this they resemble graded assessments, but the assumption is that the students will be assessed when their group completes the module, rather than when they are individually ready. Many will therefore fail (or show modest achievement); the consequence that many students will have more negative than positive reinforcement does not seem to have been faced.

Modular schemes often share many of the features of the other assessment initiatives, such as criterion-referencing and the specification of detailed objectives or assessment criteria that are independent of context. But modular schemes experience the problems of comparability in a much more acute form than other types of assessment when aggregated for the purpose of some global award like a degree or a GCSE grade. If curricular flexibility is to be encouraged, then students will present a great variety of different combinations of modules, but the assumption will usually have to be made that a grade 2, say, in one module is equivalent to a grade 2 in any other module. The more that choice is constrained (as it will be in GCSE, as argued above), the more the problem is alleviated, but the greater the loss in flexibility.

The GCSE

At the heart of the GCSE proposals is the aim of specifying curriculum content and pedagogy by a suitable specification of the examination system. (for the detailed argument see Nuttall, 1984). Of course, for any examination system to be able to succeed in this task, it must possess sufficient muscle to have its way. Historically, in Britain, there has been little doubt, even among its bitterest opponents, that the public examination system really matters to parents, students, teachers, further and higher education, and employers; that same cultural assumption has clearly been inherited by the GCSE. As we have already said, at least two of the other forms of assessment currently being developed look to the GCSE to legitimize their own status.

We should make it clear that we do not take a value position when we talk of examinations determining aspects of curriculum and pedagogy. We believe that there should be a debate on the extent to which this is desirable, and about other approaches to curriculum determination and control (described, for example, by Broadfoot, 1979). That debate is not being held now and may well be irrelevant once the GCSE is in place.

Other contributions to this volume analyse the significant features and innovations of the GCSE. Here, then, we summarize their effects upon schools, teachers and students. The aims, objectives and content of examination courses are determined (with varying precision in different subjects) by the national criteria (Department of Education and Science, 1985). Although there is provision in the criteria for their regular review and for their temporary suspension to allow for promising curricular developments to be tried, they will be difficult and time-consuming to modify. (At the time of writing, the criteria have been suspended only for one subject, SMP mathematics, a subject dear to the hearts of the top brass of the Secondary Examinations Council.) The advent of grade criteria would lead to a tightening of the specification of the curriculum. Curricular and classroom organization will be determined partly by the need to prepare students for a particular 'level' of the examination in many subjects (see Gipps's chapter above), somewhat in the present manner of choosing CSE or GCE, but possibly with a finer gradation; partly by the need for coursework and its assessment (see Macintosh's chapter); and partly by the grade criteria (see Murphy's chapter). Pedagogy stands to be influenced by the increased involvement of teachers in the process of assessment (see Torrance's chapter). Because of the requirement for comparability and the consequent need for some kind of moderation of teachers' and pupils' work, the traditional role relationship of teacher and

learner seems destined to be nudged in the direction of formal assessor and candidate, with possibly profound implications for traditional roles and with penalties for departures from the new role. We shall return to all of these core issues when discussing practical implications.

Selection and assessment

In common with existing 16-plus examinations, the principal use of the GCSE is as a selection device, whether for employers or as a requirement for training courses or advanced educational studies. The primacy of this function has particular consequences. We have mentioned already the way in which the other forms of assessment, directly or indirectly, aspire to gain a share of the action and we have alluded to the way in which this changes their role. The role of the 16-plus examination itself reflects the deeper commitment of the secondary education system to selection and much of past and present debates can be said to be concerned with the tension between the selective and other functions of education (see Broadfoot, 1979). What seems to be emerging at the end of the 1980s, however, is a willingness on the part of those concerned with devising new modes of assessment to concede that the credibility of these new modes is dependent upon accommodating to the GCSE and to the demands of the user outside the school gates — in effect, to concede the primacy of selection, albeit with extreme reluctance and distaste in some cases. It is because of the many ways in which graded assessments, modular schemes and records of achievement are having to make compromises that we feel that their initial beneficial effects within schools and classrooms cannot yet be taken as established.

In many ways, graded assessment and profiling are attempts to formalize much of the teacher-initiated informal assessment of learning which takes place as an inherent part of teaching. Such informal assessment may take a variety of guises, but its aim is formative which distinguishes it from assessment for selection, which is invariably summative. Indeed, the policy of some schools and teachers is consciously to postpone as far as possible the time at which the latter mode of assessment begins to dominate. This leads us to ask what we believe is a key question underlying all current developments, namely how far the acceptance of the supremacy of the idea of 'assessment for selection' militates against any change in the nature of teaching and learning in schools. It is not our purpose to offer speculative answers to this question, but rather to place it onto an agenda for debate. It is, of course, the striving for fairness in selection that

puts such emphasis upon comparability and reliability, and permits the down-grading of what should be pre-eminent, namely validity.

Criterion referencing

The notion of criterion-referenced assessment has acted as a principal legitimating device for the GCSE and to a lesser extent for graded assessment and records of achievement. The idea that the result of an assessment enables you to say 'what a student can do' rather than 'how a student is ranked in relation to her peers' is seen as an attractive one, and has inspired much research and debate over the past twenty years. Much of this debate has been concerned with the relationship between assessment and the curriculum, the clear specification of assessment objectives, the predictive generalizability of such assessments and the problem of defining a suitable criterion 'level' for deciding whether the 'can do' description can be awarded, and in what contexts.

It could be argued that the clear specification of assessment objectives is a desirable feature of all assessment procedures, as is an understanding of the predictive usefulness of an assessment (as one aspect of validity). In the development of graded assessments there seems to have been much concern with criterion or mastery levels, so that the probability of passing is high enough to be encouraging, but clearly not 100 per cent which would make the assessment redundant. In fact, the decision as to cut-off point is ultimately determined by considerations of student motivation, resources and so forth, rather than by some inherent logic of curriculum, pedagogy or the assessment. Hence the 'can do' skill is defined effectively in terms of the characteristics of those students who actually pass the assessment, so that to understand the meaning of 'can do' would require a detailed study of those students and the conditions under which they learnt it.

Which brings us to the issue of context. A study of the draft grade criteria for the GCSE, profile descriptors and the descriptions of levels in many graded assessment schemes shows that almost all do not specify the context in which the assessment is demonstrated. For example, one of the draft grade criteria for English in the domain of writing requires candidates to be able to 'describe a scene or character as the task requires'. There is no localization of the kind of task (perhaps the candidate can do this with some tasks and not with others), nor any attempt to elaborate on what an 'acceptable' description might be.

Such general descriptions follow inevitably from the stated aims of the assessment. Not only would context-specific descriptions be very lengthy,

but they would also be impossible to examine exhaustively. Moreover, the more specific the context, the less apparently generalizable the achievement. Since all forms of assessment are inevitably limited to a *sampling* of achievements, a high level of generality is imposed upon the descriptions. This is even more necessary in the GCSE since grade criteria have to be equally applicable to diverse syllabuses in the same subject offered by different examining groups. Yet such general, out-of-context, descriptions can only be valid on the assumption that they relate to demonstrable achievements. Thus we need to assume the existence of a high level competency, such as 'the ability to describe a scene', whose attainment an assessment system is designed to elicit. It is clear, however, that the real existence of such an ability, just like the reality of something called general intelligence, is an assumption, with important consequences. In particular, if the assumption is incorrect, then it is difficult to see what the utility of the description might be.

Suffice it to say that the degree to which achievements are genuinely context-free is very much a matter for debate. There is considerable empirical evidence challenging the notion of context-free assessment (Wolf and Silver, 1986; Nuttall, 1986). It must therefore be of some concern that graded assessments, profile descriptions and the GCSE grade criteria are all predicated on a particular view of the context-independent nature of performance descriptions.

Comparability

We have already referred to the difficulty of finding an 'objective' procedure for ensuring comparability within each form of assessment as well as between them. The experiences of the examining boards in the 1970s is salutary in this respect. They effectively abandoned the search for a foolproof procedure and settled for expert examiner judgement, in addition to a large measure of norm referencing, with all the problems of subjectivity and lack of objective reference which that implies (see Schools Council, 1979; Bardell, Forest and Shoesmith, 1978). In fact such a system resembles a cross between comparability by fiat and comparability on the basis of a strongly shared value-system. An interesting variant is the 'equivalence' between CSE Grade 1 and a GCE O-level 'pass' — an equivalence purely by fiat. No doubt something similar can be put together for the GCSE, although the increased amount of teacher assessment and the differentiated papers will make this more difficult; as a result we may see rather more equivalencing by fiat. An order of magnitude still more

difficult is the equivalencing of graded assessments and modular assessment to the GCSE. While it may be possible to specify equivalence by fiat between performance on modules and the GCSE, this will not do for graded assessments in those cases where students are assessed both by graded assessments and by the GCSE. (In other cases, the graded assessments may be accepted as the GCSE assessments.) If an equivalence is established between GCSE grades and levels of the graded assessment in a subject, then it cannot be acceptable for significant proportions of students to pass a graded assessment level and subsequently fail to achieve the equivalent GCSE grade. Yet from all that we know about the unreliability of examinations, it is almost certain that such apparent 'anomalies' will arise. This will expose both the unreliability of the assessments and the impossibility of a consistent one-to-one equivalence. Nor would it be acceptable to eliminate the problem by raising the criteria for the achievement of a given level of the graded assessment, since this will lead to even more students achieving the 'equivalent' GCSE grade but not the graded assessment level than formerly, with a perceived and presumably unacceptable fluctuation in 'standards'. Similarly, if large numbers of students also took separate modular assessments in addition to the GCSE, the same problems would arise. We note that this problem is not a new one in equating; it merely becomes apparent because large numbers of students are exposed to both 'equivalent' forms. It thus represents a major incompatibility which, if not resolved either by decoupling the equivalence or by allowing graded assessments (rigidly controlled by the national criteria) to replace GCSE, will presumably lead to the eventual discrediting of the system and perhaps to the demise of graded assessment.

Usage

There seems to be little discussion and even less evidence about the ways in which the several forms of assessment are to be used in conjunction. To a large extent the equivalencing of graded assessment and modular assessment to GCSE solves the problem (while raising others), but it remains with records of achievement. The Oxford Certificate of Educational Achievement, for example, is planned simply to be an extensive document containing public examination results as well as evidence of other cognitive and non-cognitive achievements. This presentation of detailed information is more likely to be informative than a summary, but in practice it may well be summarized into a single score or grade in different ways before being used, so that its potential utility is negated. The

importance of this can be seen within the GCSE itself where the proposals for grade criteria envisage that separate domain grades will be presented *in addition to* a summary grade over domains. This invites the use of the summary grade as the principal selection device, as occurs at present when employers and higher education selectors simply add up separate grades to form an overall score. When dealing with the more disaggregated information in records of achievement, users may tend to ignore everything other than those things which can easily be summarized into a score or grade or a picking-out of isolated and unreliable pieces of information. Here we see an incompatibility in mode of presentation of results and their likely use which, unless resolved, could lead to a downgrading of much of the information in records of achievement.

Conclusions

We have argued, in a general way, that there are both compatibilities and incompatibilities between the different forms of assessment that will determine the progress of each. We have also argued that the GCSE is the 'prestige' form to which the others are being forced to accommodate and against whose requirements their success will be judged. We have not addressed the intriguing question of whether this needs to be so, nor whether it is the result of contingencies related to sources of funding, the dominance of assessment for selection or other aspects of the educational environment. There is, however, one final issue which we feel, in the long term, is of more importance than the form of the particular kinds of assessment that survive.

All the new forms of assessment have shown at least some concern with objectives and content not previously assessed formally. Thus GCSE has raised the visibility of 'practical' assessment, and records of achievement the visibility of affective characteristics. What has not been questioned, however, is the relative status of different kinds of achievement. Put simply, our present assessment systems embody a hierarchy in which abstract, theoretical achievement is almost always accorded a higher status than the practical and the affective, and this status is intended to reflect what is believed, in the tradition of the 11-plus and 'intelligence', to influence life chances. We find little explicit in the new proposals that presents a real challenge to this view. Thus, while the GCSE pays homage to practical achievement, it is a very puny attempt to rehabilitate the status of practical knowledge and skills, and may readily be subverted. While we are witnessing some substantial changes in the *manner* of doing

assessment, we remain sceptical that they will induce any major change in the aims of education, and in its fundamental relation to the rest of society. Rather, we can easily envisage a process whereby the flagship of the assessment fleet, the GCSE, actually reinforces those aspects of present public examinations that are concerned with selection, with curriculum control and with the existing status hierarchies of kinds of knowledge.

References

- Bardell, G.S., Forrest, G.M. and Shoesmith, D.J. (1978), *Comparability in GCE*. Manchester: JMB on behalf of the GCE Boards.
- Broadfoot, P.M., (1979), *Assessment, Schools and Society*. London: Methuen.
- Bloom, B.S., Hastings, J.T. and Madaus, G.F. (1971), *Handbook on Formative and Summative Evaluation of Student Learning*. London: McGraw-Hill.
- Department of Education and Science (1984), *Records of Achievement: a statement of policy*. London: DES.
- (1985), *GCSE: the national criteria*. London: HMSO.
- Goldstein, H. (1986), 'Models for equating test scores and for studying the comparability of public examinations' in D.L. Nuttall (ed.), *Assessing Educational Achievement*. Basingstoke, Falmer.
- Harrison, A.W. (1982), *Review of Graded Tests*, Schools Council Examinations Bulletin 41. London: Methuen Educational.
- Holland, P. and Rubin, D. (eds.) (1982), *Test Equating*. New York: Academic Press.
- Inner London Education Authority (1984), *Improving Secondary Schools* (Hargreaves Report). London: ILEA.
- Nuttall, D.L. (1984), 'Doomsday or a new dawn? The prospects of a common system of examining at 16-plus' in P.M. Broadfoot (ed.), *Selection, Certification and Control*. Basingstoke: Falmer.
- (1986), 'The validity of assessments', appendix to Wolf and Silver op.cit.
- Schools Council (1979), *Standards in Public Examinations: problems and possibilities*. London: Schools Council.
- Wolf, A. and Silver, R. (1986), *Work Based Learning: trainee assessment by supervisors*. Sheffield: Manpower Services Commission.