

- Standards between Subjects, Schools Council Examinations Bulletin 29. London: Evans/Methuen Educational.
- Orr, L. & Nuttall, D.L. (1983) *Determining Standards in the Proposed Single System of Examining at 16+*. London: Schools Council.
- Pearce, J. (1983) 'The future of graded tests', *Education*, 17 June, 465-6.
- Pratley, B. (1982) 'Profiles in practice', in *Profiles*, FEU.
- Scottish Council for Research in Education (1977) *Pupils in Profile*. London: Hodder and Stoughton for SCRE.
- Scottish Vocational Preparation Unit (1982) *Assessment in Youth Training: Made-to-measure?*. Glasgow: Jordanhill College.
- Stratton, N.J. (1982a) *An Evaluation of a Basic Abilities Profiling System across a Range of Education and Training Provision*. Interim report for CGLI Profiling Project 3. London: CGLI.
- Stratton, N.J. (1982b) *Reliability of basic skills profiles*. Paper given at the British Education Research Association 7th Annual Conference, St. Andrews.
- Willmott, A.S. and Nuttall, D.L. (1975) *The Reliability of Examinations at 16+*. London: Macmillan Education.

## 13 Profiles and Graded Tests: the Technical Issues<sup>1</sup>

DESMOND NUTTALL AND  
HARVEY GOLDSTEIN

*Some of the technical issues which Baumgart identifies briefly in Chapter 4 are here explored in considerable detail. Very few people have written anything about the technical problems that at least some of the approaches of profiling and records of achievement raise. Yet it is on these issues that the success of this innovation may well ultimately depend, for if the information in the records cannot readily be relied upon or used it will have no public credibility, and the whole edifice that so many people have been at pains to build in the last decade or so will crumble as completely as previous efforts to initiate such records have done. Whilst for many such minutiae may seem tiresome in comparison with the educational arguments involved, Nuttall and Goldstein demonstrate in this chapter that their resolution is an essential prerequisite to any system of recording achievement that aspires to more than a local focus.*

—Editor

We shall review first, but briefly, what seems to be the current state of profiling; the aims and controversies which are being discussed. We

<sup>1</sup> This article is reproduced with the kind permission of the Further Education Unit who first published it in 'Profiles in Action', FEU, 1984.

shall then identify what seem to us to be the key technical issues which have yet to be solved before profiles can be accepted widely as satisfactory additions or alternatives to other forms of assessment. Then we shall describe and evaluate the graded test movement. Finally we shall link together graded testing and profile reporting and discuss the relationship between them. On all these topics we have been able to find little written about technical matters, and very little relevant empirical research, and in this article we therefore discuss the major technical issues in the hope of facilitating a more informed debate.

### CURRENT ACTIVITY

The last few years have seen a considerable interest in profile reporting in education. An early report from the Scottish Council for Research in Education (SCRE, 1977) has stimulated a variety of others both at the school level in the work of the Schools Council (Balogh, 1982)—and even more vigorously at the FE level in the work of the Further Education Unit (FEU, 1982a and b), and the related City and Guilds of London Institute (CGLI) development work (Stratton, 1982a).

There is general agreement that the disaggregated nature of the information in a profile is of greater potential use to students, lecturers and employers who wish to understand the specific strengths and weaknesses of an individual, rather than some average assessment. There is also a general recognition of the need to distinguish formative profiles which are produced during a course from summative ones which represent a final assessment and are typically based on an amalgamation over time of the formative set of assessments. Formative profiles are designed to be part of the learning process, to be discussed by lecturer and student together, while the summative kind are designed for the outside world and selection for employment. These two broad purposes may be often in conflict and affect the need for comparability and reliability in the assessments, as we shall demonstrate.

There is also conflict among the motives of the advocates of profiles. At one extreme, in the words of Broadfoot (1982), herself a collaborator in the SCRE work, profiles provide a viable alternative to the powerful 'anti-educational' constraint of public examinations. At the other is the Associated Examining Board's new A-level Geography examination which provides a profile by means of a separate grade for

each of four papers, to yield more information from existing examinations, a development reviewed in detail by Harrison (1983). In between, in the mainstream of development, is the FEU and CGLI approach and that of SCRE and many of the schools studied in the Schools Council work, which see their profiles as existing alongside centralised examinations. In the case of the pre-vocational courses stimulated by the FEU, the intention seems to be that profiles should become a dominant element in assessment, while in schools there seems to be a recognition that public examinations will remain the dominant element for some time to come.

One feature common to most profiles, in contrast to examinations, is their inclusion of personal and social skills alongside so-called basic skills and more conventional attainments. Sometimes the personal and social skills are reported simply as records of activity, without judgement, but often they are assessed and graded (by the lecturer alone, or by the lecturer and student in partnership). In the CGLI Vocational Preparation course (365), for example, there are four ratings for each of the abilities 'to be self aware' and 'to cope with problems', as well as for 'calculating' and 'reading and writing'. Simple rating scales of this kind are widely used; each point is accompanied by a description of the ability or attainment level it represents. While it is usually admitted that some of these ratings will be more 'subjective' than others, the common aim appears to be to remove this element as far as possible by providing careful verbal descriptions, training the assessors carefully and carrying out periodic 'quality control' checks on their work. Some more open-ended schemes use 'comment banks', consisting of comments whose placing on the scale has had prior agreement (Black and Dockrell, 1981).

This present article is largely concerned with some of the key technical issues which have received too little serious attention in the discussion up to the present. Such references as there are admit to the existence of problems of reliability and indeed validity, but fail to probe them in any depth (Macintosh, 1982; Harrison, 1983). In our view, however, there do exist serious technical problems and, unless these can be solved satisfactorily, profiles will rest on insecure foundations. We are concerned largely with the 'mainstream' developments, although we do recognise other strands to profiling.

## NORM VERSUS CRITERION-REFERENCING

This article does not provide the space to present a detailed discussion of this issue, but some general remarks are relevant. This is because, for both profiles and graded tests, criterion-referenced assessment is commonly advocated. Further, since criterion-referenced assessment has often been regarded as not amenable to quantitative manipulation in the same way as norm-referenced assessment, there is the possibility that some important technical problems will be ignored.

Firstly, it is extremely difficult to imagine a criterion-referenced assessment that is totally independent of norm levels. If a criterion point is to be useful it must obviously distinguish between individuals, so that the possibility exists that some will reach it and others will not. We can know whether this is the case only by collecting data on how many individuals do so—thus estimating a norm. Since, in fact, 'cut off' points are a choice, if not actually arbitrary, this 'norming' information will typically be used in determining them, at least initially. The difference between criterion-referenced and norm-referenced tests lies in their methods of construction, intended use and interpretation. In terms of construction, criterion-referenced tests are designed consciously to avoid the psychometric models of norm-referenced tests, and in terms of use they employ fewer categories but ones which are designed to convey educationally meaningful information. In the case of profiles, the important issue is that of providing assessments that are both accurately related to the profile elements and comparable across individuals. This is not easy to achieve.

A central thrust of the profiles movement has been the attempt to relate assessment more closely to the curriculum than public examinations typically do (Mansell, 1982). Thus, teachers and lecturers have collaborated in the design and testing process, and assessments are intended to be made in the context of curriculum activities. If curricula differ, a comparability problem is immediately raised, since there is then no guarantee that consistent interpretations can be made. To overcome this problem, much effort has been put into an attempt to develop context-free assessments. The SCORE profile, for example, has one skill level for number which is described as 'Can handle routine calculations with practice', and the amplification is provided 'Fairly accurate but slow'; is able to calculate percentages and money calculations etc.

The difficulty with such out-of-context descriptions is that they are too poorly defined to ensure comparability, and the more precisely defined they become the more rooted in a context they become. Thus

the above definition would need to specify what was meant by 'routine' so that such a calculation could be recognised. It would need to specify the order of difficulty of the 'percentages' referred to and make much more precise the phrase 'Fairly accurate but slow', and so on. Eventually, for a high degree of comparability to be achieved, the description would have to be so precise that we would be very nearly back with the classical test situation where everyone is administered effectively the same set of items—i.e. a highly specific context for the assessment. Of course such tests need not be paper-and-pencil ones. They could be imaginative practical or work-related assessments, but still be contextual, and thus would encounter the same problems as all traditional tests; namely not being equally relevant to each of a wide diversity of curricula.

It seems clear, therefore, that there is an inbuilt contradiction here that is not only unresolved, but also hardly discussed in the recent literature. Moreover, given the difficulty, if not the sheer impossibility, of achieving comparability between existing public examination boards because of the differences between the syllabuses and courses it is difficult to see a more satisfactory solution emerging for profiles (Goldstein, 1982).

If profiles are to be faithful to a curriculum, then they will presumably have to sacrifice the aim of comparability across curricula, and cease striving to become context-free. This raises the possibility that some (key) elements of the profiles will be 'centralised' and others 'localised'. In the circumstances it seems very likely that the former will come to assume greater importance than the latter. It is interesting to note that government policy on the 17+ is that in 'key' areas (English, science, maths) performance will be 'externally assessed or moderated' while at the same time the policy generally supports profiles (DES, 1982a). Whether the assessments are norm-referenced or criterion-referenced is secondary.

Before leaving this topic, it is worth pointing out that the skill descriptions used in profiles so far developed for the school or further education system presuppose, because of the context-free requirement, that there really are abilities or skills which can be applied equally within different contexts. Thus, in mathematics, skills are defined in terms of symbolic mathematical operations so that a child who can 'calculate a percentage', for example, presumably can do so in all practical contexts. What many researchers have realised is that such symbolically defined skills do not necessarily transfer from one situation to another, since performance depends upon disposition and motivation, for example, as well as competence, and indeed that the autonomous existence of a 'skill' is itself rather a slippery notion. It is

arguable that this issue is fundamental to skill assessment and that if profiles ignore it their relevance will be greatly diminished.

This discussion has been mainly in terms of comparability, which is an attribute of assessment principally required by the selector (for employment, further or higher education) who wants to compare one individual with another. With formative profiles, comparison with others is less important; the record can then concentrate upon what the student has done, and under what conditions, so that the context is specified. The record can be cumulative, so that development of skills or their application in new contexts can be observed and discussed: the comparison is with the individual's own past and not with others' current performance. Reliance upon grades and grids is no longer necessary, and some schemes like the Record of Personal Experience (see de Groot, Chapter 7) consist only of the students' own accounts of their experiences (validated by an adult), without comment or judgement by the teacher.

In those schemes where a common framework is needed for the recording of judgements, valid, reliable and comparable assessment can be achieved only if assessors are trained to interpret the concepts used. Discussions of what is meant by the concept, the contexts or occasions in which it might manifest itself, and the evidence that would allow judgements to be made about the degree to which it is present, are all essential. Discussion followed by ratings of particular examples can generate a calibrated comment bank that will allow others to be trained more quickly in the use of a particular scale. Nevertheless, as pointed out above, once assessments are required to reflect learning contexts, true comparability becomes elusive. This will almost certainly happen if individual teachers or lecturers rate their own students.

There have been few studies of the degree to which the different scales in the profile are assessing separate attributes. In Scotland it was found that teachers were reluctant and often unable to make distinct assessments of personal qualities like honesty (SCORE, 1977), while Stratton found that improvements on one attribute were almost invariably matched by improvements on another (e.g. 'working with authority' and 'self awareness') though his findings were not consistent across his two samples of raters (Stratton, 1982a). While more studies of this kind would be useful, one must accept that lack of discrimination between attributes may arise through the 'halo' effect, that is, the tendency of a rater's overall impression (favourable or unfavourable) of the student to influence all the individual ratings. More training of raters, especially more discussion of specific examples of behaviour, may well reduce the 'halo' effect and so convey more real information.

## SCALING OF PROFILE ELEMENTS

Whilst not all profile systems scale, if an employer or a teacher is offered a profile in which a number of disparate elements are each rated on a scale of 1-4 say, there is a clear invitation to compare the ratings. A student might be rated 1 (high) on 'planning a task' and also 1 on 'working with colleagues'. Yet if the former is applicable only to say 5 per cent of the population and the latter to 25 per cent it is difficult to see how the two ratings can be equated. Indeed, the only method of satisfactorily equating them is to define the ratings as applying to the same population percentages. Thus this would imply a direct 'norm-referencing' of the grades, and is again an issue which has been very little discussed. It is relevant to note the public examination boards' experience with equating grades in different subject areas. After some early work in the 1970s (Nuttall et al., 1974) the attempt was abandoned once it was realised that this could be done only on a strictly norm-referenced basis as described above, thus violating the principle that the grading systems should not be solely norm-referenced.

There seems to be no good reason why there should be the same number of steps for each element. The number should be determined through experience and discussion so that the lowest should be in reach of all, and the highest, neither a ceiling which all reach nor a level which few reach. In some cases there may be room for several intermediate steps, in others for one or even none. The Scottish Vocational Preparation Unit has also drawn attention to the giant and uneven steps in many schemes, which could be avoided if the definitions of the scale points were generated by experienced practitioners familiar with the range of attainment in the population and using real examples of student behaviour.

Scaling done in this manner, leading to an appropriate number of steps for each element rather than forced into a uniform mould, might be less prone to invite inappropriate comparisons between ratings while, at the same time, leading to more attainable goals for students and lecturers.

## WEIGHTING AND COMBINING PROFILE ELEMENTS

The grid type of profile tends implicitly to give equal weight to each element, leaving the user to choose the subset on which she or he

wishes to concentrate. Once this subset is selected, however, there is still an implied equal weighting so that the user, in the absence of specific guidelines, presumably will attach equal weight to the selected elements. Yet, for a variety of reasons this may be inappropriate. Some elements may be measured with low reliability (see later), some skills may effectively appear several times in slightly different guises, some assessments may have stronger validity than others, etc. In other words, the user generally needs more information about the profile other than the profile itself, just as the traditional test user should have access to information on reliability, validity, norms, etc. Yet, given the already large quantity of data supplied in some profile systems, the provision of such extra information seems somewhat daunting. Some research to study users' needs and the way in which they are used the information supplied would be welcome.

The weighting problem becomes of crucial importance if a user is to aggregate all or a subset of the elements. Not only will the above considerations apply, but the user will have her or his own relative weights and some guidance would be useful. In the absence of such guidance, there is a danger that many users will, often inappropriately, average in some simple fashion the ratings, grades or scores.

Weighting and combining elements is particularly tempting if performance is recorded quantitatively, and the temptation to make inappropriate combinations might be less in schemes where numbers are not attached to the descriptions of behaviour or evidence. The temptation to weight and combine elements is also reduced if each element does not have the same number of scale points.

## RELIABILITY

Quite a lot has been written about the reliability of grading systems, especially in public examinations (Wilmott and Nuttall, 1975) and it is now widely recognised that quite large measurement errors exist, so that there is a reasonably high probability that a student with a particular grade could have obtained a grade one or even two removed on a parallel examination, for example, one with a different set of questions or with a different marker. The reliability of the elements of a typical profile, often assessed subjectively or perhaps by means of a short skills test, could be very low, much lower than that of a public examination. Yet there is a negligible amount of serious effort devoted to studying this problem. Of course, as Macintosh says, validity is

fundamental and we have already said something about that. However, if a very unreliable profile is interpreted too literally by a user, serious mistakes can occur. Consider, for example, the SLAPONS profile designed to communicate arithmetic skills to employers (Prattley, 1982). Each element has a 'score' of from 0 to 5, yet as with many conventional examinations, there is little indication of whether a difference of 1 or 2 or 3 score points between students or between elements is to be treated as meaningful or could be within 'measurement error'.

It is, of course, quite difficult to obtain estimates of measurement error (the standard error of measurement as it is known in the context of standardised tests) and the most popular traditional methods seem of little use (Ecob and Goldstein, 1983). The measurement errors can arise from a number of sources. There are differences between assessors or raters. There is a variation in the tasks on which students are judged and there is variation in the response given by the student from day to day or situation to situation. Also, in a profile, some of these measurement errors may be correlated and their effects thus compounded.

Stratton studied the agreement between raters by asking them to place examples of behaviour on the profile scale (Stratton, 1982b). There was consensus for 71 per cent of the examples, though for about a third of these the consensus may have been spurious. A sound consensus therefore emerged only with just less than half the examples. The raters were, however, inexperienced, and Stratton concluded that agreement might be much higher among trained, experienced raters. But this study shows the magnitude of error that may arise from just one source, and reinforces the need for considerable careful research in order to provide some indication of measurement error. For example, a set of confidence intervals, based on rough estimates, one for each element, could be devised so that judgement of differences would occur only for non-overlapping intervals. These could also, in principle, be incorporated visually onto a profile chart, as is often provided with standardised test batteries, and we would suggest that those who are preparing profiles pay particular attention to this possibility.

Drawing attention to measurement error is particularly important with summative profiles because of the importance of the decisions that might be made in the light of the information contained in them. With formative profiles, where irrevocable decisions can be avoided, lower reliability might be tolerated if an increase in reliability can only be achieved at the expense of validity. But it is likely that the techniques used to enhance reliability, like more training, the use of

two or more raters or gathering more evidence, are also those that will enhance validity by promoting clarification and deeper understanding of each element in the profile. Thus we again return to the importance of collaborative development and operation of profile systems, in which training occurs through discussion and rating of examples, and where the possibilities and limitations of a profile system can be illuminated.

### WHITHER PROFILES?

We have, quite deliberately, emphasised the current technical shortcomings of profile research and implementation. We do so not because we wish to argue against profiling as such, in fact quite the contrary because we believe that profiles do have interesting potential. It is because we are concerned that a too ready acceptance of a technically weak system will ultimately be counter-productive when its deficiencies become apparent during use. As we have indicated, in the well-established area of public examinations there are still considerable technical problems to overcome and in the sophisticated area of statistical test theory and psychometrics these controversies over fundamentals continue to rage. In both these areas, part of the case against the assessment techniques has rested on technical inadequacies. We are quite clear that the technical problems surrounding profiles are just as difficult as in these other areas and to ignore them would seem to be folly.

In our view, it would be wise to spend time now reflecting on these technical matters before too widespread and too rigid systems are developed. From a research point of view there is no doubt that there are considerable challenges, and in the areas of reliability, scaling, weighting and studying 'skills' it should be possible to make useful progress.

### GRADED TESTS

The graded tests movement shares many of the aims of the profiling movement, for example, a desire that education and assessment are seen as positive rather than negative experiences for all students, and a

determination to put the curriculum first. Well-established in sport, music and other performing arts, graded tests are relatively recent arrivals in mainstream subjects of the secondary school curriculum, but have already made a dramatic impact upon the teaching and learning of modern languages and, in the Kent/Schools Council Mathematics Project, upon mathematics.

As Baumgart argues in Chapter 4, the basic idea of graded tests is not new and might be considered part of normal good practice. Phase tests in Technician Education Council (TEC) units, and indeed the TEC system of units at progressively higher levels (e.g. Maths 1, Maths 2, Maths 3), are straightforward examples of graded tests, where progress to the next level is contingent upon success at the previous level (a success that comes to most, if not all, students).

Yet graded tests have suddenly begun to attract a good deal of attention. The Cockcroft Report (DES, 1982b) has given graded tests—called 'graduated' tests there—further respectability and, in response to its recommendations, the DES has announced a substantial programme of research and development on graded tests in mathematics, principally for low attainers. Some of the modern language schemes are also designed principally for low attainers, but others are for the full ability range, as are most of the schemes in sport and the performing arts.

Perhaps the best known scheme (and certainly the oldest, founded some 100 years ago) is run by the Associated Board of The Royal School of Music. It attracted nearly 350,000 entries from the UK and Eire in 1980, an average of over 50,000 for each of the first five grades and sharply fewer (below 20,000) for the top three grades which involve a theory component as well as a practical. Each of the grades is designed to represent a defined standard of performance while the grades together form a progressive sequence of development in practical musicianship. The examination can be taken several times a year and the grades are not tied to particular ages, so that the scheme is tailored to the progress of each individual. Furthermore, as with sports, the choice of test items or pieces tends to be limited, with many elements, such as scales, known in advance.

Similar features, apart from the last, are characteristic of virtually all graded test schemes. In his review of graded tests Harrison (1982) encapsulates the essence of a typical graded test scheme in modern languages in three features: 'that it is progressive, with short-term objectives leading on from one to the next; that it is task-oriented, relating to the use of language for practical purposes; and that it is closely linked into the learning process, with pupils or students taking the tests when they are ready to pass.'

The curricular side of the schemes is especially significant in modern languages, where the movement is known as Graded Objectives in Modern Languages (GOML) to emphasise that the guiding principle is in the development of a well-defined progression of educational objectives (building from the bottom upwards) rather than in the tests themselves.

Advocates of GOML schemes (of which there were about 60 in 1981, according to Harrison) point to the increased motivation of their pupils, with tangible proof given by dramatic increases in the proportion of third-year pupils opting for modern language courses in their curriculum in the vital fourth and fifth years. They also report that pupils and parents value the certificates issued for each grade.

Nevertheless, in most of what has been written about graded tests (as about profiles) there is little dispassionate evaluation, and it is therefore difficult to analyse what the key ingredients of their success in motivating pupils really are. One is almost certainly the short-term nature of the objectives allied to the principle of mastery learning and testing by way of criterion-referenced tests, which more than 90 per cent will pass in modern languages (and more than 80 per cent in music): that is, positive reinforcement, coupled with a tangible reward, at relatively frequent intervals (in contrast to public examinations where reward is stored until the end of five years of secondary education, and then is granted to the few rather than the many).

Another key ingredient is the enthusiasm of the teachers, sparked off, it would seem, by the curricular innovations of GOML. GOML schemes tend to be local and the teachers using the schemes have the chance to be involved in the further development of the schemes and in the assessment of their own pupils. This professional commitment is reminiscent of the early days of the Certificate of Secondary Education (CSE) examination, and may well be dissipated (as many would say has happened in the CSE) as the schemes become routine or are taken up by those who have not been party to the original development. In Oxfordshire, HM Inspectorate judged that in many schools there was still far to go in thinking about appropriate objectives for the less able, which stresses how difficult it is to translate laudable aims into effective classroom practice (HMI, 1983).

The most common type of problem raised by GOML teachers in the survey conducted by Harrison was organisational, for example, a lack of secretarial help, additional demands on time or the organising of the oral tests themselves. But it is apparent that the organisational problems posed by individual rates of progress have largely been ducked by teachers, who have used the expedient of testing all pupils at about the same time. HMI are critical of this failure to meet what most would

regard as one of the cardinal principles of mastery learning:

'On the whole, however, it [the use of Oxfordshire's graded test scheme] tends to be confined to the less able pupils who all take the test at the same time. . . . Even where groups have this measure of homogeneity it is clear that some pupils are being faced with too easy a test which for others is still too difficult. . . . Thus, while the original intention of a high pass-rate is achieved, the timing of testing is more often related to the age of the pupils than to their linguistic readiness.'

Sport and the performing arts tend to be taught either individually or in small groups not as closely linked to age as school year groups, and therefore avoid the problem of modern languages. But that problem is likely to be as acute in mathematics and other subjects for which graded tests are currently being proposed, namely science and English.

It is clear that the individualisation of learning, something that is likely to be accelerated by the microprocessor, creates organisational problems for an individual teacher or lecturer within the normal institutional constraints. Those concerned with whole institutions, such as principals, heads and LEAs, need to consider the implications of the widespread introduction of graded tests, and at least one Chief Education Officer has already welcomed the implied break with the tradition of grouping pupils by age and the possibility of grouping by attainment level or developmental stage instead (Brighouse, 1982).

#### THE CURRICULUM AND GRADED TESTS

Mansell remarked that 'profiling forces assessment into the learning process' (Mansell, 1982); the worry of many is that graded tests will force assessment into the learning process not in the constructive way hoped for by Mansell, but in a destructive way that will lead to an excess of testing and to a backwash effect throughout the secondary school or further education curriculum that will make the backwash effect of the much maligned systems of examinations at 16+ look mild in comparison. Curriculum-led assessment may be a splendid concept when the agreed curriculum commands enthusiasm and support, as it manifestly does in modern languages, but where there is no agreed curriculum or where the field of study is so vast that several different curricula are equally acceptable (as is the case of English), the particular subject curriculum that comes to lead the graded tests may be viewed by many teachers as a straitjacket. On the other hand,

Pearce sees graded testing acting as a stimulus to sort out some of the problems of curriculum diversity:

Graduated testing on any scale would expose the flaws in our position with painful clarity. The real need is a machinery to enable teachers to negotiate agreed curricula and institutionalize those agreements with a necessary minimum of validation. That is what has happened with BEC, and to a different extent with TEC, with on the whole very encouraging consequences as well as the inevitable protests of those in whom the loss of their chains induces a state of terror' (Pearce, 1983).

One way to guard against the worst sort of backwash is to ensure that the assessment procedures validly measure the full diversity of curricular objectives. That requirement almost certainly demands an impressive array of oral, practical and written assessments, as well as course work, projects and other extended exercises so that we should talk of graded assessments rather than graded tests. Couple these assessments with the need for at least two formal occasions of testing each year at each grade level, and one arrives at a substantial assessment industry that is viable only if the teachers themselves accept a major role in the assessment of their own pupils.

The technical issues in assessment of pupils by teachers have been studied extensively (for example, Cohen and Deale, 1977) and can be summed up in the two concepts of reliability and comparability. With criterion-referenced graded tests, achieving agreement about the criteria for marking among all those involved might be simpler than it is within traditional public examinations, but the variation in the conditions under which the tests are given and the variation in the tasks from school to school, and occasion to occasion, may wipe out any enhanced reliability of marking. Since there is a ready opportunity to retake a graded test, it might be argued that reliability of assessment is not as important as it is within the public examination system but this will depend upon the significance of the decisions made as a consequence of the test result.

There is a further fundamental difficulty which arises particularly acutely when a test can be retaken, namely the opportunity to learn, or teach to, the test itself, so that the curriculum will become distorted. To avoid this problem, tests would have to be changed between administrations and a moment's reflection will indicate the enormity of the task of continually developing new tests and equating them with the old, where testing takes place two or three times a year. The investment of time and expertise that this process requires is well represented in Holland and Rubin (1982). We are not aware that this issue has been faced seriously by the advocates of graded tests.

If mastery is in fact essential before a student can successfully work

at the next level, then a false positive (a pass given when a fail should have been) may be as damaging as a false negative, which denies a student who is ready to move up the opportunity to do so. The consequences for individual students, especially at higher levels, where the results are more obviously for external consumption, may lead to too great an emphasis on striving for high reliability. As with public examinations, the fear would then be that the demand for high reliability will override the demand for validity. This tension appears to be a common feature of assessment systems, and the direction in which it is resolved tends to be a function of the significance of the decisions made on the test results. Open entry to higher education, for example as in the Open University, would reduce the significance of A-level grades.

Similar concerns arise with the pressure for comparability. The more standardisation that is imposed in the quest for reliability and comparability, the greater the threat to the key features of graded tests. One area in which comparability might reasonably be sought is over the number of levels in a graded test scheme designed principally for the age range 11 to 16. Most of the modern language schemes have five levels, and the Cockerott Report suggests between four and six. But in the first case, the five levels span the full ability range and five years of secondary education, while in the second the target is just low attainers from the age of 14 upwards.

#### DETERMINING GRADE LEVELS

What considerations are important in the choice of the number of levels and their positioning or spacing? Educational theorists are not in sufficient agreement in most fields to provide an answer, and so the choice will be guided largely by practical considerations such as balancing the value of frequent feedback to students with the desire to avoid excessive testing. How the grades of the graded tests might be linked to the grades of public examinations, most obviously those at 16+, if indeed such a link is either desirable or feasible, is also a matter for much discussion in the GOML movement as more and more schemes develop level 4 and 5 material (for more detail about these issues, see Harrison, 1982 and Kingdom, 1983).

Of more fundamental concern, as Baumgart also argues in Chapter 4, is whether the concept of progression from one grade to the next makes sense in many subjects of the curriculum. While almost all



subjects are taught on a broad principle of progress, this progress is not tied to the linear development of an unvarying set of objectives and there are many different ways of progressing through the same syllabus. Mastery of the objectives at one level may, therefore, not be essential to the study of the objectives at the next level, and graded tests could easily become simply modular tests, that is, tests on self-contained content that can be taken in any order and whose material can be forgotten without apparent penalty after the test has been taken.

In practice, the lack of differentiation and individualisation in education (and the Oxfordshire modern language schemes are probably typical in this respect) probably serves to make the graded part of GOML tests relatively insignificant: the important things are teacher enthusiasm and pupil rewards. A modular scheme might serve just as well. The most precious ingredient, therefore, becomes teacher enthusiasm, which puts a premium on local self-determination and involvement and argues against making national or regional comparability so important that the development of graded tests becomes simply another centralised assessment activity.

Another issue is whether attainment of a grade or level, particularly when specified in criterion-referenced terms, can satisfactorily be determined at the gross level of a subject or has to be at a much more disaggregated level as we have discussed earlier in the case of profiles. In public examinations Orr and Nuttall (1983) argue that true criterion-referencing and aggregation are incompatible, and the same arguments would seem to apply to graded assessments. Harrison draws attention to the uneasy compromise between global certification and criterion-referencing that seems to be arising in some of the GOML schemes. Dealing with more narrowly defined skills or domains may help to make the progression through the grades more obvious, and allows for some skills to be put into cold storage at some levels while new ones are introduced, thus adding more flexibility in those cases where there is no single route of progress. At the same time, reasonably reliable separate assessments of many skills at each level may magnify the testing load unbearably, especially for the assessor.

## PROFILES AND GRADED ASSESSMENTS

Thus there are a number of common, or very similar, issues facing

profiles and graded assessments. In particular, the choice of the elements or dimensions that should be assessed deserves much deeper thought and investigation. At present, the dimensions have been chosen for sound educational (curriculum-led) reasons but without much subsequent exploration of overlap and redundancy. Sometimes a single dimension embraces multiple objectives that are better separated. More careful specification of the objectives and the evidence needed to determine whether they have been successfully or partially achieved would clearly be beneficial, and could be followed after the event by the straightforward analyses used by Stratton (1982a) to detect redundancy (or possible 'halo' effect).

Deciding upon the number of reporting levels and the size of the steps is also a shared problem. Its solution must be rooted in the experience of teachers and lecturers whose knowledge of the typical performance and the range of performance in the particular population of pupils or students is vital. But too great a reliance on the norms of the past should be avoided; both profiles and graded assessments have stimulated unexpected improvements in motivation and attainment, and the definition of the steps should therefore be carried out in action rather than determined in advance.

This leads to the suggestion that more needs to be established about the effects of profiles and graded assessments upon students and lecturers. It was suggested above that the notion of 'grades' or 'progression' might be relatively unimportant and that the key ingredients were public rewards for the students and the enthusiasm of the teachers, but this is still speculative. Investigations of profile schemes in action (following Coacher, and Stratton) and evaluation of the new graded test developments are essential. This is particularly important where graded test schemes are being devised only for the 'low attainers', since the dangers of labelling in this procedure are only too obvious; there are, in any case, considerable but rarely discussed problems in actually defining, for example, the 'lowest 40 per cent of the ability range'.

Both profiles and graded tests make the curriculum that leads them much more obtrusive than the system of conventional examinations. The consequence is that, if lecturers and teachers are to preserve their freedom in choice of teaching strategy and examples relevant to the local context, both kinds of developments must be local rather than national (though a national framework is, of course, not ruled out). The inherent limitations in the concept of comparability must be exposed so that the advantages of formative assessment can be permitted to flourish. But by putting emphasis on the local, the pressing need for training of assessors is also emphasised and made more urgent.

accent in assessment moves to stress its formative value rather than its summative use, so the research should develop so as to be sensitive to such changes.

## REFERENCES

- Balogh, J. (1982) *Profile Reports for School Leavers*. York: Longman for Schools Council.
- Black, H.D. & Dockrell, W.B. (1981) *Diagnostic Assessment in Secondary Schools*. London: Hodder and Stoughton for SCRE.
- Broadfoot, P. (1982) 'The pros and cons of profiles', *Forum*, 24, pp. 66-9.
- Brighouse, T. (1982) *Education*, 24/31 December, p. 491.
- Burgess, T. & Adams, E. (ed.) (1980) *Outcomes of Education*. London: Macmillan Educational.
- Cohen, L. & Deale, R.N. (1977) *Assessment by Teachers in Examinations at 16+*. Schools Council Examinations Bulletin 37. London: Evans/Methuen Educational.
- Department of Education and Science (1982a) *17+ : A New Qualification*. London: HMSO.
- Department of Education and Science (1982b) *Mathematics Counts* (the Cockcroft Report). London: HMSO.
- Ecob, R. & Goldstein, H. (1983) 'Instrumental variable methods for the estimation of test score reliability', *Journal of Educational Statistics*, 8, 3.
- Further Education Curriculum Review and Development Unit (1982a) *A Basis for Choice*. London: FEU.
- Further Education Curriculum Review and Development Unit (1982b) *Profiles*. London: FEU.
- Goacher, B. (1983) *Recording Achievement at 16+*. York: Longman for Schools Council.
- Goldstein, H. (1982) 'Models for equating test scores and for studying the comparability of public examinations', *Educational Analysis*, 4(3), 107-18.
- Harrison, A.W. (1982) *Review of Graded Tests*, Schools Council Examinations Bulletin 41. London: Methuen Educational.
- Harrison, A.W. (1983) *Profile Reporting of Examination Results*, Schools Council Examinations Bulletin 43. London: Methuen Educational.
- Her Majesty's Inspectorate (1983) *A Survey of the Use of Graded Tests of Defined Objectives and their Effect on the Teaching and Learning of Modern Languages in the County of Oxfordshire*. London: DES.
- Holland, P. & Rubin, D. (ed.) (1982) *Test Equating*. New York: Academic Press.
- Kingdom, J.M. (1983) *Graded tests*. Paper presented at an informal seminar at the University of London Institute of Education.
- Macintosh, H.G. (1982) 'A 17+ package: a view from the school', in *Profiles*, FEU.
- Mansell, J. (1982) 'A burst of interest', in *Profiles*, FEU.
- Nuttall, D.L., Backhouse, J. and Willmott, A.S. (1974) *Comparability of*

Although some profile schemes imply that each dimension is judged in the same way on a four-point scale, the evidence for the judgements can be of very different kinds, ranging from single subjective appraisals of personal qualities to cumulative test-based assessments of numerical skills. In the latter case, there is potentially a very obvious marriage of profiles and graded assessments.

It comes as no surprise, then, that profiles and graded tests are being brought together. The Cockcroft Report envisaged that performance in the graded tests might contribute to the kind of profile described above, an idea that has rapidly been taken up by a number of LEAs and examining bodies. For example, ILEA are proposing a 'London Record of Achievement', a portfolio containing details of examination passes, other achievements in school and a profile compiled by teachers, parents and the pupils themselves (see Mortimore and Keane, Chapter 5). The portfolio will also contain the results of graded tests in mathematics, English, a foreign language (European or Asian), science, and design and technology, though development work has begun only in mathematics, English and science.

Even more advanced are the plans for the Oxford Certificate of Educational Achievement (OCEA) which Willmott describes in Chapter 9, that will link a profile with graded assessments and examination results on one certificate.

The principal contrast between graded assessments and profiles (anyway, in their record form rather than the grid form) lies in their stance towards quantification and measurement. Graded assessments are firmly within the psychometric tradition of tests and examinations, while the advocates of profiles are often against measurement and the reductionism and trivialisation that all too often accompany measurement. So, in the union of profiles and graded assessments, we see the exciting prospect of bringing together the humanistic and quantitative traditions in educational assessment. From this could emerge a most fruitful collaboration that could give a new rigour to humanistic assessment while preserving the pre-eminence of validity and curricular relevance.

To achieve this, action research is essential, integrating the development with the evaluation, and analysing the processes of selection, judgement and interpretation in the development and use of the assessments within the context of the college, school and workplace. While the technical issues we have discussed also need to be studied, what is not needed is the sort of detailed technical research of the type that has been done for 50 years on existing examination systems, largely atheoretical and motivated by a desire to provide merely technical answers to essentially educational problems. Just as the