Profiles and Graded Tests: the Technical Issues

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

1 This article is reproduced with the kind permission of the Further Education Unit who first published it in "Profiles in Action", FEU, 1984.
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) in 1979 recommended that profile reporting should be introduced at the school level in the work of the Schools Council for Educational Creativity (SCEC) and even more vigorously in the FE level in the work of the Further Education Council for Scotland (FECOS). There is general agreement that the disaggregated nature of the information contained in profiles is likely to give teachers new insights into their unstructured teaching, and even more complete insights into the weaknesses of an individual. Rather than some vague assessment, teachers often want to know something specific and quantifiable. The new profile is designed to be a part of the learning process, to be representative of a full final assessment, and to be a result of the formative assessment which represent final comment and not simply a stand-alone grade, or even a grade which does not exist. A profile is a full formative assessment, a full representation of the learning process, and the scores are designed to be the result of the formative assessment, and not just a final grade.
The 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 profiling and guided tests, criterion-referenced assessments are commonly advocated. Further, since criterion-referenced assessment has been referred to as not amenable to quantitative judgement 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—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, interpretation, and in terms of use. Criterion-referenced tests are designed to assess individual profiles only, whereas 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 relax assessment more closely to the curriculum than previous examination methods try to 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. It differs, 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 common sets of procedures. The SCRE profile, for example, has one skill level for number which is described as 'Can handle routine calculations with ease', and the amplification is provided 'Fairly accurate but slow, is able to calculate percentages and money calculations etc.

The difficulty with both out of context descriptions is that they are too poorly defined to ensure comparability, and 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—that is, 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 context-bound, and thus would encounter the same problems in 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 unevaluated, but also rarely 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 those (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 crave striving to become context-free. This raises the possibility that some 'key' elements of the profiles will be 'centralised' and others 'localised'. In the communalities it seems very likely that the former will come to possess greater importance than the latter. It is interesting to note that government policy on the 17's is that in the 'key areas' (English, science, maths) performance will be 'externally assessed or moderated' while at the same time the policy generally supports profiles (DES, 1982). 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, without different contexts. Thus, in mathematics, skills are defined in terms of symbolic mathematical operation 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
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 sense if the user devotes 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 (Pratley, 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 (Eccle 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, 1982). 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 profiling and implementation. We do 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—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 curriculum and graded tests

Mandated national content standards and their associated assessments are a core component of every state’s education system. These standards are developed by state education agencies or councils and are based on the content areas that students must master in order to graduate from high school. They are designed to ensure that all students have the knowledge and skills necessary to succeed in post-secondary education and their future careers.

The standards are typically divided into academic subjects such as math, science, English, history, and social studies. They specify what students should know and be able to do at each grade level, from kindergarten through grade 12.

Graded tests, such as standardized tests and end-of-course exams, are used to measure how well students have mastered the content standards. These tests are administered at regular intervals throughout a student’s academic career and are used to assess student progress, inform instruction, and identify areas where additional support is needed.

The results of these tests can impact student placement, eligibility for scholarships, and even college admissions. High schools may also use the results of these tests to determine whether students have met graduation requirements.

Overall, the curriculum and graded tests play a crucial role in the education system, ensuring that students are prepared for the challenges of the future and that schools and districts are held accountable for their educational outcomes.

104 Profiles and records of achievement

The primary role of the school is to ensure that all students have the opportunity to achieve their full potential. Therefore, it is essential to establish a clear and well-defined curriculum that provides a solid foundation for learning. This includes the development of a coherent and sequential curriculum that is aligned with the state content standards and benchmarks.

Achievement is best measured through a combination of formative and summative assessments. Formative assessments, such as quizzes and projects, provide ongoing feedback to both teachers and students, while summative assessments, such as quarterly exams and end-of-course exams, measure overall learning outcomes.

The importance of achievement cannot be overstated. It is essential for students to develop a strong understanding of the material being taught, and for teachers to provide the necessary support to help students achieve success. By establishing a clear and well-defined curriculum and regularly assessing student progress, schools can ensure that all students have the opportunity to achieve their full potential.

105 Profiles and records of achievement

The primary role of the school is to ensure that all students have the opportunity to achieve their full potential. Therefore, it is essential to establish a clear and well-defined curriculum that provides a solid foundation for learning. This includes the development of a coherent and sequential curriculum that is aligned with the state content standards and benchmarks.

Achievement is best measured through a combination of formative and summative assessments. Formative assessments, such as quizzes and projects, provide ongoing feedback to both teachers and students, while summative assessments, such as quarterly exams and end-of-course exams, measure overall learning outcomes.

The importance of achievement cannot be overstated. It is essential for students to develop a strong understanding of the material being taught, and for teachers to provide the necessary support to help students achieve success. By establishing a clear and well-defined curriculum and regularly assessing student progress, schools can ensure that all students have the opportunity to achieve their full potential.
DETERMINING GRADE LEVELS

What considerations are important in the choice of levels and their positioning or spacing? Educators and others are not always agreeable to this, yet the decision must be made, and the decision must be made in a way that is fair and consistent. The choice will be made in terms of the value of feedback, the need for the results to be used in decision-making, and the need for the results to be used in decision-making. The choice will be made in terms of the value of feedback, the need for the results to be used in decision-making, and the need for the results to be used in decision-making. The choice will be made in terms of the value of feedback, the need for the results to be used in decision-making, and the need for the results to be used in decision-making. The choice will be made in terms of the value of feedback, the need for the results to be used in decision-making, and the need for the results to be used in decision-making.
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 encompasses 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 Stratto (1982a) to detect redundancy (or possible 'hula' 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 grades 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 draw 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 in 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 unnecessarily, especially for the assessor.

**Profiles and Graded Assessments**

This there are a number of common, or very similar, issues facing...
Profiles and graded tests

Although some profile evidence is generally available, the individual student plays a very different role in the use of profile evidence in deciding on the profile, and the evidence is not intended to be a substitute for more traditional forms of evidence.

It seems likely that, in practice, the student's evidence will be used in the systematic and graded assessments of the student's performance in the examination. The student will, for example, be expected to complete a self-assessment of their profile evidence, and this will be used in the systematic and graded assessments of the student's performance in the examination.

In the case of the student's evidence, the evidence is not intended to be a substitute for more traditional forms of evidence, and the evidence is not intended to be a substitute for more traditional forms of evidence.