# *Graded Assessment and Learning Hierarchies in Mathematics*

**RICHARD NOSS, HARVEY GOLDSTEIN & CELIA HOYLES,**
*Institute of Education, University of London*

ABSTRACT *The Graded Assessment in Mathematics (GAIM) project has sought legitimation, at least in part, by appealing to notions of child-centred learning. At the same time, in order to gain acceptance, it has attached itself to a certification procedure, namely the General Certificate of Secondary Education (GCSE). This paper argues that the demands imposed by the needs of certification are irreconcilable with a requirement for child-focused 'diagnostic' style assessment, and have led to serious internal contradictions. The paper also offers a critique of some of the research concerned with establishing hierarchies of mathematics learning.*

> Isn't it interesting how often we find ourselves using the idea of level? We talk about a person's levels of aspiration or accomplishment. We talk about levels of abstraction, levels of management, levels of detail. Is there anything in common to all the level-things people talk about? Yes: they each appear to reflect some way to organise ideas—and each seems vaguely hierarchical. Usually, we tend to think that each of those hierarchies illustrates some kind of order that exists in the world. But frequently those orderings come from the mind and merely appear to belong to the world. (Minsky, 1986, p. 91)

## Introduction

From the early 1980s in the UK there has been considerable discussion of assessment in schools, and a growth in the allocation of resources invested in developing assessment schemes (see, e.g., Goldstein & Nuttall, 1987). Much of this interest can be traced to Central Government involvement, first with the Assessment of Performance Unit (APU, 1986) and more recently with its investments in schemes to develop graded assessments and records of achievement (Nuttall & Broadfoot, 1987), and its creation of the Secondary Exams Council (SEC) to oversee school leaving exams. The 1987 Education Bill placed before Parliament envisages an expansion of the SEC to co-ordinate the regular testing of children at various ages between 7 and 16 years.

Within this concern with testing—largely for the purpose of school and Local Education Authority accountability—existing assessment schemes and developments of them will undoubtedly form a part of the new pattern. In our view, there exists a fundamental inconsistency in these initiatives based on a confusion between two distinct objectives of assessment. Indeed in order to clarify this confusion, it might be worth attempting to reserve the word *assessment* for one objective alone: the development of an awareness, through diagnosis, of children's progress in order to aid learning. As such, assessment is a necessary part of teaching. If it is accepted that teaching mathematics is more than merely the transmission of knowledge, and that the pupil, in order to learn mathematics, has to construct her own understandings on the basis of her existing schemas, then teaching must of necessity take account of pupil conceptualisations. It seems to us that this kind of assessment—which is essentially private between the teacher and the pupil, and which can loosely be termed 'diagnostic'—is a process in which every good teacher is involved; its aim is to promote learning by virtue of a detailed study of a pupil's strengths and weaknesses.

In contrast, we would wish to designate as *grading* the process by which children's achievements are evaluated for the purpose of sorting and selection for employment or educational opportunity. This kind of grading (of which public examinations are exemplars) is essentially public, i.e. the results are intended to grade or rank students or institutions, and to report such grades or ranks so that others can form judgements about pupils (and, perhaps, their teachers).

Of course, it would be inappropriate to posit a firm separation between public and private assessments, in that, for example, public assessments can be used for subsequent curriculum development. The problem is that schemes which purport to *simultaneously* undertake both grading and diagnostic assessment are fundamentally unviable. The danger exists that schemes which start out solely with the intention of providing private or diagnostic tools become used as part of public assessment—and thus their original intention is destroyed. Our specific concern in this paper is with these schemes—which claim to diagnose, but whose implicit or explicit function is grading. It is our view that this applies in the field of mathematics with schemes such as the Graded Assessment in Mathematics (GAIM), and we show below why this is so.

## CSMS and Hierarchies in Mathematics

The Cockroft report (DES, 1982) has been extremely influential in the development of policy in relation to mathematics education in the UK. The findings of the Concepts in Secondary Mathematics and Science study (CSMS; Hart, 1980), had a very evident influence on the report, particularly those sections which related to assessment. The CSMS study claims to provide evidence for the existence of specific hierarchies of attainment in mathematics. On close inspection, however, these claims exhibit serious weaknesses.

CSMS looked at a wide range of topic areas [1]. Within each topic area they studied 13-, 14- and 15-year-olds, a total of several thousand children being tested in over 50 schools. The main analysis can be described simply as follows:

(1) After some preliminary analysis, items were 'clustered' together within diffi-

culty bands for each topic. That is, for items with roughly the same level of difficulty, those items which were strongly associated among themselves formed a 'group'.

(2) After some further, relatively minor, rearrangement the groups were ordered into 'levels' such that if a child 'passed' one level he or she would probably have passed lower ones. Children were deemed to have 'passed' a level if two-thirds of the items in the group were correct.

(3) These levels are essentially the hierarchies which the project set out to establish.

The first point to notice about this analysis is that the grouping of items by difficulty and strength of association necessarily implies their subsequent scalability as in (2), although the level of associations found did not allow a complete scalability. Furthermore, items which did not 'fit', i.e. were not strongly associated with a group of other items, were omitted from the analysis. In other words the 'hierarchies' consisted of those items, reflecting particular aspects of the topics, which happened to fit into this hierarchial pattern. In essence, therefore, the report presents us with groups of items which are of about the same difficulty and where (relatively) success on any one implies success on any other in the group.

The CSMS report, however, goes beyond the derivation of a set of groups or levels, and places a particular interpretation upon them: namely that within each topic they represent an actual hierarchy of *understanding,* although it is acknowledged that this interpretation is "opinion rather than fact" (Hart, 1981, p. 208). In fact there is no detailed discussion as to how such an interpretaton is obtained. Initially CSMS used the term 'hierarchy' in the Piagetian sense of a series of stages which represent some kind of 'natural' pattern of progress in learning. While the study subsequently rejected a strict Piagetian framework for the selection and classification of test items, it appears that the initial assumption of the existence of actual hierarchies of understanding led the CSMS team to interpret their levels on the basis of such a theoretical stance.

Underlying the CSMS interpretation there seems to be a view of the existence of an ordered set of abilities among children. This view has been commented upon by Ruthven (1986) who characterises the "ability stereotyping" evident in recent government reports such as *Better Schools* and *Mathematics Counts,* as "... the ordered acquisition of a clearly defined and structured corpus of knowledge and skill, mediated by the cognitive capability of the individual pupil" (p. 41). Here we have a view of mathematics as an ordered hierarchy, and of pupils' mathematical abilities as correspondingly hierarchical. This view is very evident in the discussion of the CSMS findings in the section titled 'Implications for teaching': "It is *impossible* to present abstract mathematics to all types of children [*sic*] and expect them to get something out of it" (Hart, 1981, p. 210).

The view of mathematical knowledge described above is not universally accepted, and nor therefore are interpretations of research studies of mathematical attainment which we see as arising from this view. It is arguable whether mathematics possesses a hierarchial structure in the sense that a notion at any given level is defined uniquely in terms of lower level ideas. Of course hierarchies do exist on a local level—clearly there are interconnections and chains of definitions and argument within and across the network of concepts which make up the mathematical domain. This does not tell us anything however about the global nature of the subject, still less about the psychological and pedagogical implications of such relationships. As Vergnaud points out, such hierarchies of understanding as do exist are in reality at lower, more primitive levels than competency in a given mathematical topic:

> ... the hierarchy of mathematical competencies does not follow a total order organization, as the theory of stages unfortunately suggests, but rather a partial order one: situations and problems that students master progressively, procedures and symbolic representations they use, from the age of 2 or 3 up to adulthood and professional training, are better described by a partial-order scheme in which one finds competences that do not rely on each other, although they may all require a set of more primitive competencies, and all be required for a set of more complex ones. (Vergnaud, 1983, p. 4)

We would maintain that there are other theoretical interpretations of the CSMS data which are consistent with the results, and which could explain its findings, provided that there is not an *a priori* assumption of the existence of a hierarchy of understanding. An important omission from the research in our view is any relevant data on the details of topic exposure. Since there were a large number of schools involved, it is reasonable to assume that there was differential exposure between schools to the different topics. The report itself partly recognises this. We know from other research (for example the Second International Mathematics Study (Goldstein, 1987, ch. 5) that performance is influenced by topic exposure. Thus some, if not most, of the observed associations among items may simply reflect the *differentiation in topic coverage* between schools.

An alternative interpretation which is, in our view, equally plausible is that the observed hierarchies reflect not the differences but an underlying homogeneity in the mathematics curricula—that is *similarity in topic coverage* between schools. In this scenario, the hierarchies would reflect what actually happens in schools, rather than any universal levels of understanding. They say nothing therefore as to what might be achieved given a different set of experiences. Put another way, it is important to remember that a study concerned solely with observed patterns of performance of itself can say little about alternative patterns in response, for example, to new curricula or teaching methods.

It should be clear that neither of the interpretations above is inconsistent with the CSMS findings, and that neither requires the existence of any single hierarchy of understanding which is invariant for all children. When implications are drawn for the learning of mathematics in school, the issue becomes even more problematic. There is no obvious connection between levels of understanding (even if they do exist) and sequences of learning. It seems to us that the important issue is not to seek some kind of invariant sequence through which children inevitably pass, but to map the variety of possible sequences which might be viable. Furthermore, we should entertain the possibility that no one sequence may be optimal, and that some sequences work better with some children. In fact, a note of warning is sounded by one of the CSMS researchers, who argued in the context of learning algebra, that "it might seem sensible to base the teaching given to children at levels 1 and 2 on the meanings for the letters that these children readily understand. On closer examination this is by no means a straightforward task. For example, the use of letters as objects totally conflicts with the eventual aim of using letters to represent numbers of objects . . . " (Küchemann, 1981, p. 119).

There are two problems associated with the development of an assessment scheme based on hierarchies. Firstly, such a scheme is likely to be interpreted in the school context as a recipe for curricular sequencing. School mathematics is, at the present time, largely compatible with a step-by-step hierarchical view since it fulfils characteristics dictated by the educational discourse. It is important to be aware of the epistemological and psychological distortions produced by this 'didactical transposition', and to find ways to challenge rather than reinforce them. Secondly, once in place, such hierarchies can become self-fulfilling—so that those who seek to base curriculum innovation on a belief in mathematical hierarchy are locked into a tautological loop.

In order to focus our ideas, we now turn to one of the assessment projects in mathematics—Graded Assessment in Mathematics (GAIM)—which has specifically made use of the CSMS results to build an assessment scheme in this way.

## GAIM

The GAIM project has been in the process of development since 1983 by a team of teachers and researchers. The GAIM team have produced a variety of modes of assessment which include open-ended tasks, practical activities and group as well as individual work. Some important and unique features of the GAIM assessment scheme are that it is based solely on course work without a terminal component, it provides pupil profiles and it leads to external certification if desired. Many of the activities designed by GAIM are interesting, original and provide an excellent basis for curriculum development.

GAIM is an assessment scheme for 11- to 16-year-olds which is designed for use alongside any secondary mathematics curriculum. It has a structure of 15 levels, the top seven being designed to be equivalent to GCSE grades. Levels arise from topic criteria which are a bank of so-called 'criterion-referenced' ('can do') statements which describe the mathematics which students 'know, understand and apply'. The topic criteria are organised into six areas (logic, measurement, number, space, statistics, algebra and functions) "which cover both process and content". The key notion within GAIM is that of a *level*, which we shall discuss from two perspectives: within a topic area, and between topic areas.

### *Levels Within a Topic*

The overall hierarchy within a topic area—the Within-topic criteria in GAIM—started from a consideration of research studies such as CSMS and the APU. They have then been developed in consultation with teachers from 10 local education authorities who made the decisions about the difficulties of items or the levels which could be ascribed to different pupil responses based on their collective experience. As with the CSMS study itself, the teachers' decisions will, of course, reflect current curriculum practice.

The individual topic criteria consit of a series of 'can-do' statements. This is immediately problematic in that if such statements do provide a framework for diagnosis in any meaningful sense they would have to be very specific: this would necessitate a massive number of statements. We have serious reservations about the notion that a pupil can or cannot do a given task in a clear-cut way. In the first place, the notion that a task can/cannot be done presupposes that there is no room for performance to be gauged in terms of partial success—it assumes that performance is all-or-nothing. A similar point can be made in relation to understanding: current approaches in psychology are stressing the fragmented nature of understanding (see, for example, Minsky, 1986) and in particular how quite elementary understandings cannot be thought of as a straightforward can/cannot choice. More fundamentally, we wish to question the whole notion of *context-free* assessment. There are two main contextual influences on performance which any notion of assessment needs to take into account: the *setting* within which the activity takes place, and the *conceptual tools* available. We consider each of these in turn.

The APU research contains a number of items which measure performance on the same mathematical operation embedded in different settings and have shown quite dramatically different facility levels. If, for example, we embed mathematical questions in a money context, pupils are more likely to obtain correct answers than when exactly the same questions (from a mathematical point of view) are presented differently. For example, at age 11, the item

$$4.5+0.5=$$

was answered correctly by 63% of pupils. In contrast, despite the 'word-problem' setting, the question:

> John saved £3.70 and then his mother gave him £1.50. How much did he have in all?

was answered correctly by 82% (APU, 1986, p. 836).

These differences in facility are provoked by merely embedding mathematical operations differently within word problems or presenting them 'purely' symbolically. The difficulty is further compounded when we turn to the assessment of *practical* activities on which many of the GAIM tasks are based, since the way people solve mathematical problems depends crucially on the discourse in which the problem is seen to be situated. There is evidence of a gap between intuitive mathematics displayed, for example, in everyday settings, and that exhibited in school situations. Lave and her colleagues (Lave *et al.*, 1984) have shown that calculations which shoppers in a supermarket performed were almost 100% correct, but when asked to perform the identical calculations with paper and pencil, performance fell to about two-thirds of that in the supermarket setting. Other evidence indicating the role of cultural factors in influencing mathematical attainment is available in the work of Carraher *et al.* (1985). The point is that there will almost certainly be crucial differences in strategies and facilities which vary according to the pupils' definition of the situation—for example, as a mathematical task or a shopping task.

The second major contextual influence we wish to consider is the nature of the tools available to pupils, and their effect on what children can and cannot do. These points are well illustrated by the following example of a GAIM topic criterion:

> Can a) sketch a cuboid (with or without isometric paper) and b) make it by first drawing its net. (SPACE Level 7; GAIM, 1987)

As far as we are aware, no research evidence exists (either from the GAIM team or elsewhere) to suggest that the presence or absence of isometric paper in the task of drawing a cuboid has no effect on facility level. On the contrary, we would suggest that isometric paper can be thought of as a conceptual tool for the solution of this problem, in a manner similar to that of a calculator in arithmetic—and when looked at in this way must surely be a factor influencing performance. The evidence from studies of computer-based mathematical environments (such as Logo; Hoyles & Noss 1989; Noss & Hoyles, 1988) indicate that the computational tools available crucially influence both the way pupils work and what they can achieve.

As well as these major influences on performance, there are a range of other well-researched influences on performance levels. For example, the APU has shown that the phrasing of a question can be critical: "'How many halves are there in $2\frac{1}{2}$?' has a success rate which is 30% more than '$2\frac{1}{2} \div \frac{1}{2}$' at age 11, and 29% more at age 15. Using 'of' instead of a multiplication sign when fractions are multiplied also substantially improves performance at both ages" (APU, 1986, pp. 840–841). The APU study similarly illustrates that the introduction of diagrams or the use of non-standard conventions can influence performance. Nesher & Tecibal (1975) show that the use of verbal cues can improve performance (such as 'more' means 'add') or lower performance (when 'more' does *not* mean 'add'). Unfamiliar words, whether they be important for the mathematical solution (such as 'product' or 'isosceles') or even when they are not (such as questions related to 'overs' in cricket) can make questions more difficult. Similarly, the way in which test items are administered will also affect performance (Murphy, 1980).

The fact that the 'same' mathematics can be harder or easier depending on the context and manner of presentation is by no means a new discovery, and not in itself crucial to the validity of a given test, as long as relative difficulty between items is recognised. However this issue becomes extremely important if one is attempting to build criterion-referenced topic hierarchies. Notions of decontextualised 'can-do' statements must be strictly meaningless in any criterion-referenced sense. All statements about hierarchies much be conditional on stated contexts and the available tools—they cannot be universal. We propose that rather than searching for 'context-free' statements about achievement we should instead be attempting to describe the different settings in which assessment can take place, and observing the differences and similarities which arise.

### *Equating Levels Across Topics*

The need to have the same number of levels for each topic in GAIM which implies, for example, that 'a level 7 task' has a certain meaning across topics, arises from the requirement for assigning particular GAIM levels to GCSE grades. We argue that both the equivalence of GAIM levels to a more traditional terminal GCSE, and the consequent equivalencing of levels across topics, have no theoretical rationale. It is hard to imagine a framework which would provide a basis for equivalencing two tasks allegedly at a given level from disparate areas of the mathematics curriculum—areas which constitute essentially different domains of knowledge. In our view the claim for criterion-referencing within a particular knowledge domain must necessarily rule out the possibility of equivalence on the basis of any theoretical justification. Thus the only possibility for legitimising such equivalencing is to consider the issue from an empirical (norm-referenced) point of view.

It has long been recognised in the area of public examinations that attempts to equivalence across exam subject or topic areas in the case of such terminal public exams have no theoretical support (Wood, 1977). The examination boards have been concerned with this issue for a long time and have experimented with a variety of techniques in an attempt to ensure comparability of grades across different boards and across years, within a given subject. Forrest & Shoesmith (1985) give an account of current procedures, and point out that "the study of comparability is by no means a straightforward activity". Indeed, as Goldstein (1986) points out, there are both theoretical and practical arguments which suggest that a satisfactory comparability between grades is unobtainable. We shall not here rehearse these arguments, save to point out that the only situation where two or more assessments may be validly equated is where each one legitimately can be substituted for any other one for a given group of examinees, without altering the outcome in terms of ranking.

We have already pointed out that GAIM, as with other graded assessments, consists of simple 'can/cannot do' judgements. Thus, to ensure comparability between judgements, we have to consider the following scenario. Suppose we can ascertain, for the pupils in school A, their probabilities of fulfilling a specified criterion at a specified level at a specified time, using a given assessment procedure, with a particular teacher or a particular task. Likewise, suppose we can ascertain the probabilities for the same level for the pupils in school B where, inevitably, a different assessment procedure is used. To equate the assessments, a typical procedure would be to give a large sample of pupils from each school both assessments and then plot the results of one assessment against the other. If, for the sake of argument, we suppose that the raw assessment is made on a pseudo-continuous scale, then it is clear that complete equatability will be obtained when a monotonic relationship between the assessment results exists. If it does not then we can obtain inconsistencies such that the order relationship implied by one assessment is different from that implied by the other. In practice, of course, this will nearly always occur, so that we will have to form a judgement of the 'acceptable' degree of equatability (Goldstein, 1986). The issue is therefore one which needs to be studied empirically.

Thus, in the light of what has been said of the conditions for equating to take place, to equate GAIM results we should need to set up a procedure, say for written components, as follows: A sample of pupils taking a more conventional terminal GCSE would also be given the written tasks appropriate to the levels of GAIM to which we wanted to assign the GCSE grades. Let us say that we wished to assign grades C, D, E of GCSE to levels 13, 12, 11 as is proposed by GAIM. Suppose also we found that equating was possible when results were plotted against each other. Even so, such a study would not be sufficient to demonstrate equatability between the terminal GCSE and GAIM levels, since the GAIM tasks were undertaken earlier and at different times during a pupil's school career. To ensure such equatability we would need to assess the same children on the GAIM levels during their school careers and then also when they took a terminal GCSE. Not only does this imply a complex and difficult, if not impossible, series of studies, it also implies that we could not even demonstrate that equating was possible until *after* those pupils had taken a terminal GCSE. If it were the case that a satisfactory equating had not been achieved, then for that cohort of children, in retrospect, the equivalence would have to be abandoned, and even when a satisfactory equating had been found, there would still need to be a continual monitoring to check whether it was maintained. The resource implications of this, and the consequences for those pupils who had retrospective changes of judgement made, would seem to rule this out of court.

If, instead of such 'statistical' equating, an examiner-based 'moderation' procedure was used, the difficulties are no less. In essence examiners would need to carry out similar procedures, albeit more subjectively, to those of a statistical procedure, and it is not at all clear how they could do this. As we have pointed out, the learning of any subject, including mathematics, does not take place in a strictly hierarchical fashion, in the sense that we can say for any pupil that understanding of a given set of concepts necessarily implies understanding of a specifiable 'lower level' set. Learning can validly follow different paths for different pupils, as can teaching. By the end of a course, the set of understandings possessed by any given pupils will be a cumulative function of those obtained during the course. But two students may have the same set of final understandings while having different sets at any given stage of the course, and so at least having the possibility of being assessed differently by a single graded assessment while the course is in progress. Likewise, two students may exhibit the same set of understandings at a given stage of the course while finishing with a different set of understandings. In these circumstances a unique equating between the assessment attached to a level and that attached to a final assessment is in general not possible.

Everything which has been said about written assessments applies to other forms of assessment. We have argued that to be fair, and to be seen to be plausible, GAIM as well as other graded assessment schemes must, at the very least, go to considerable trouble and expense to equate levels across topics, across institutions and across time. Furthermore, if they wish to assign GCSE grades they would have to get involved in even more complex studies which, on purely theoretical grounds, can be shown to have doubtful validity. As we see it, this does not necessarily constitute an argument for the abandonment of GAIM. There may well be much valuable material that has been and could be developed which will find curriculum uses and be of help in assessment. What does seem to us worth abandoning is the attempt to incorporate GAIM into the GCSE, and thus to make it an assessment that really counts in the sense of affecting life chances in a very direct way. It should be noted that this applies with equal force to any future attempts to link GAIM or GAIM-style grading techniques to assessments at, say, 7, 11 and 14 as is currently proposed.

**Assessment-driven Curriculum Change**

Curriculum change tends to be a slow process and it is often suggested that an effective way of accelerating such change is to introduce an innovative assessment procedure. We wish to briefly consider this claim in terms of its practical effect on classroom practice, from a number of different perspectives: the effect on the pupils' motivation, on the mathematical activies themselves, and on the relationship between teacher and pupil.

*Motivation.* One of the claims of the GAIM team (and of a number of other schemes—not all of which are based on similarly investigatory material), is that the introduction of assessment as an integral and continuous component of children's mathematical activities, will lead to an increase in pupils' motivation. There is, we suggest, little *a priori* evidence that this is the case: indeed, there is some evidence that the introduction of extrinsic rewards can actually lead to a decrease in intrinsic motivation (Deci, 1972, 1975).

*The effect on mathematical activities.* An issue which concerns all attempts to lead a curriculum via assessment, is the danger of narrowing curriculum content. In the case of GAIM, this revolves around the specific hierarchies themselves. It might be very tempting, when confronted with a list of topic criteria, to teach to those criteria alone. While the danger of this might be mitigated by requiring teachers to undertake investigative work, it is possible that this will be seen merely as 'doing investigations' rather than adopting an investigative approach to doing mathematics. It is our view that if investigations are assessed then they will tend to change their nature. The essence of an investigation is its openness and the extent to which it provides a vehicle for exploration. It is difficult to see how grading—i.e. an assessment *linked to a certification*—will encourage this. Thus, criteria for the assessment of investigations implicity or explicitly will be agreed between teachers and pupils. The aim will be to satisfy the agreed criteria rather than to encourage a wide ranging creativity.

*The pupil–teacher relationship.* One result of an increased emphasis on assessment is an extra burden placed on teachers. GAIM itself emphasises the importance of teachers spending large amounts of time conducting assessments, and indeed, have argued that the linking of GAIM levels to GCSE assessments "... is evidently required, to avoid duplication of effort on behalf of teachers and students ..." (Brown *et al.*, 1987). It is relevant to ask, therefore, whether the assessments, given the overall pressures on teachers' time, will become highly formalised and routine in order to keep them manageable.

A more fundamental point concerns the effect of graded assessment on the relationships within the classroom. There have been moves in the UK towards a 'negotiation' of mathematical meanings between pupil and teacher, the adoption of a more exploratory approach, and a readiness on the part of many teachers to see participation in (and enjoyment of) mathematical activity as a primary goal in the classroom. As more and more of the normal classroom activity becomes the subject of certified assessment, it will become important to 'score' well, at least in the pupils' view, and the learning support role of the teacher may become subtly altered in the direction of an external assessor. We should not underestimate the extent to which the grading function of testing can actually run counter to its diagnostic function: an emphasis on grading can all too easily encourage the suppression of pupils' weaknesses and a concentration on maximising assessment ratings or test scores, rather than aiding in the diagnosis and remediation of misconceptions.

## Conclusion

It has been a painstaking process, but slowly over the past two decades some teachers have found convincing ways of making maths enjoyable and of valuing —at least until the public examination race begins—what children can actually do, rather than counting how many mathematical hoops they can jump through. Once hoops are put in place, they tend to become fixed—*irrespective* of changes around them, such as those brought about by technological innovation, or new insights into how children learn.

We do not wish to be misunderstood on the reasons for choosing GAIM as our target: we have levelled our criticism at the GAIM initiative precisely because it exemplifies the problem of the confusion of roles for 'assessment'; it has made a genuine attempt to tackle the question of assessment, rather than simply concentrate on grading, and it is the relationship between these two roles which we have sought to clarify. We certainly are not advocating the imposition of tests which avoid this confusion by being completely oblivious to the diagnostic objectives of testing (in the name of 'accountability') and focusing entirely on grading: such developments seem to us to be entirely concerned with the desire to 'appraise' teachers and to publish league tables of schools—in our view they offer nothing of value to the educational effort.

We have argued, however, that programmes which pose questions of sequencing in terms of universal necessary stages, have no theoretical or empirical underpinning; they also provide an unfortunate model for schools and teachers to adopt. Accepting the notion of a simple hierarchy (which *must* be the case if GCSE grade assignment is to take place in the final instance), and claiming its legitimacy, allows others to adopt notions of universal 'benchmarks' representing 'progression' through schooling, and hence gains credence for the next step—that is, standardised competency testing at selected ages.

In conclusion, the first report of the Government's Task Group on Assessment and Testing (DES, 1988) recommends establishing sequences of levels in each assessment component, very much of the type we have discussed in this paper. Hence, the criticisms we have made and the changes we have discussed should be taken into account in any consideration of this proposed system of national assessment targets. In our view, rather than expend resources on developing systems of attainment targets, it would be educationally more valuable to point to their essential unviability, particularly where there has been no serious evaluaton; and to sound a warning that these schemes may become extremely damaging both to individuals, or groups of individuals, as well as to the process of experimentation and innovation in education.

*Correspondence:* Dr Richard Noss, Department of Mathematics, Statistics and Computing, Institute of Education, University of London, 20 Bedford Way, London WC1H 0AL, United Kingdom.

## NOTE

[1] These were: measurement, number operations, place value and decimals, fractions, positive and negative numbers, reatio and proportion, algebra, graphs, reflections and rotations, vectors and matrices.

## REFERENCES

ASSESSMENT OF PERFORMANCE UNIT (APU) (1986) *A Review of Monitoring in Mathematics, 1978–1982* (London, Department of Education and Science).

BROWN, M., BURNS, S. & ONION, A. (1987) Going for GAIM, *Times Educational Supplement,* 8 May.

CARRAHER, D., CARRAHER T. & SCHLIEMANN, A. (1985) Mathematics in the streets and in the schools, *British Journal of Developmental Psychology*, 3, pp. 21–29.

DES (1982) *Mathematics Counts*, Report of the Committee of Enquiry into the Teaching of Mathematics in School (London, HMSO).

DES (1988) *Task Group on Assessment and Testing* (London, Department of Education and Science).

DECI, E. L. (1972) Effects of externally mediated rewards on intrinsic motivation, *Journal of Personality and Social Psychology*, 22, pp. 113–120.

DECI, E. L. (1975) *Intrinsic Motivation* (New York, Plenum Press).

FORREST, G. M. & SHOESMITH, D. J. (1985) *A Second Review of GCE Comparability Studies* (Manchester, Joint Matriculation Board).

GAIM (1987) *Graded Assessment in Mathematics Development Pack: teacher's handbook* (Kings College, University of London).

GOLDSTEIN, H. (1986) Models for equating test scores and for studying the comparability of public examinations, in: D. L. NUTTALL (Ed.) *Assessing Educational Achievement* (London, Falmer Press).

GOLDSTEIN, G. (1987) *Multilevel Models in Educational and Social Research* (London, Griffin; New York, Oxford University Press).

GOLDSTEIN, H. & NUTTALL, D. (1987) Can graded assessments, records of achievement, and modular assessment co-exist with GCSE?, in: C. GIPPS (Ed.) *The GCSE: an uncommon examination*, Bedford Way paper No. 29 (Institute of Education, University of London).

HART, K. (1980) *Secondary School Children's Understanding of Mathematics*, Report of the Mathematics Component of the Concepts in Secondary Mathematics and Science Programme (Chelsea College, University of London).

HART, K. (Ed.) (1981) *Children's Understanding of Mathematics 11–16* (London, J. Murray).

HOYLES, C. & NOSS, R. (1989) The computer as a catalyst in children's proportion strategies, *Journal of Mathematical Behavior*, 7, pp. 53–75.

KÜCHEMANN, D. (1981) Algebra, in: K. HART (Ed.) *Children's Understanding of Mathematics 11–16* (London, J. Murray).

LAVE, J., MURTAUGH, M. & DE LA ROCHA, O. (1984) The dialectic of arithmetic in grocery shopping, in: B. ROGOFF & J. LAVE (Eds) *Everyday Cognition* (Cambridge, Mass., Harvard University Press).

MINSKY, M. (1986) *The Society of Mind* (New York, Simon & Schuster).

MURPHY, R. (1980) Sex differences in GCE examination entry statistics and success rates, *Educational Studies*, 6, pp. 169–178.

NESHER, P. & TECIBAL, E. (1975) Verbal cues as an interfering factor in verbal problem solving, *Educational Studies in Mathematics*, 6, p. 41.

NOSS, R. & HOYLES, C. (1988) The computer as a mediating influence in the development of pupils' understanding of variable, *European Journal of Psychology of Education*, 3, p. 3.

NUTTALL, D. & BROADFOOT, P. (1987) *Pilot Records of Achievement in Schools Evaluation*, Interim Evaluation Report (Open University & Bristol University).

RUTHVEN, K. (1986) Differentiation in mathematics: a critique of *Mathematics Counts* and *Better Schools*, *Cambridge Journal of Education*, 16(1), pp. 41–45.

VERGNAUD, G. (1983) Why is an epistemological perspective a necessity for research in mathematics education? In: *Proceedings of the Conference of the North American Chapter of the International Group for the Psychology of Mathematics Education*, Montréal, 1983.

WOOD, R. (1977) Your chemistry equals my French, *Times Educational Supplement*, 20 May.