

## Criterion-referenced assessment

A common example of a simple criterion-referenced examination is a driving test. Here a dichotomous view of criterion-referencing is adopted wherein the driver either can or cannot perform the elements of the test – reversing round a corner without mounting the pavement, performing an emergency stop without ejecting the examiner through the windscreen, overtaking without killing passers-by. The example of a driving test is useful for it indicates a problem in criterion-referenced assessment, *viz.* whether a student needs to pass each and every element (to meet every single criterion) or whether high success in one element can compensate for failure in another. Strictly speaking in an overall criterion-referenced system the former should apply (Gipps 1994), which is draconian since a student could fail a whole course by dint of failing one small element, a common feature of the driving test. This is a problem that has been recognised in modular courses, particularly where one or more compulsory modules have to be passed (Morrison 1993). This also raises another problem in the educational sphere, *viz.* that the dichotomous view of pass/fail – *either* able *or* unable to perform a task, is unhelpful to teaching and assessment. Abilities are not of the ‘either/or’ type where one can or cannot do a task; rather, a task can be performed with differential degrees of success in its different elements or as a whole.

Another example of a criterion-referenced assessment is the music performance examinations of the Associated Board of the Royal Colleges of Music. Let us say that I am taking one of their piano examinations that requires me to play the scale of C sharp minor in double thirds for four octaves. If I play it successfully I pass that element of the examination; if I play it very well I receive a higher mark in that element of the examination; if I play it to the standard of a concert pianist I receive an even higher mark. Then I fail my performance of a Bach fugue, missing notes, failing to bring out the fugue’s subject in the stretto sections and playing the wrong notes, thereby failing this element of the examination. However, the music examination differs from the driving test in that my marks in the former are *aggregated* so that one strong element can compensate for one weak element, so that, overall I might pass the music examination even though I am carrying a failed element.

The example of the piano playing examination is useful also, for it indicates that a measure of subtlety can be used in handling marks. These stand in contrast to the National Curriculum assessments where *levels* not marks are being used; this is problematic because marks are more reliable than levels, retaining a degree of detail that is automatically lost when marks are combined to give a level. The problem is exacerbated further in the National Curriculum because the levels that were crude aggregates of marks are themselves aggregated to give an overall level in a subject. We are only one stage away from aggregating all of the subjects to give a single level index of a student! There are several advantages in using *marks* rather than *levels* for assessment purposes:

- (i) it enables partial completion of a task to be recognised, wherein students can gain marks in proportion to how much of the task they have completed successfully;
- (ii) it affords the student the opportunity to compensate for doing badly in some elements of a task by doing well in other elements of a task;

- (iii) it enables credit to be given for successful completion of parts of a task, reflecting considerations of the length of the task, the time required to complete it, and the perceived difficulty of the task;
- (iv) it enables weightings to be given to different elements of the task and to be made explicit to the students;
- (v) scores can be aggregated and converted into grades.

The question of aggregation is troublesome for criterion-referenced tests (Gipps 1994). If one were being true to the specificity of criterion-referencing then this would argue against aggregation at all, as aggregation – by collapsing details in an overall aggregate – loses the very specificity and diagnostic/formative potential upon which criterion-referencing is premised. However if one fails one criterion out of a large number of criteria in an assessment then one fails the overall assessment.

The problem of aggregation in criterion-referencing is compounded because criterion-referencing suggests the compilation of long lists of criteria for each element of an assessment, echoing the dangers of behavioural objectives (Morrison and Ridley 1988). The parallel with behavioural objectives is not idle, for, just as with behavioural objectives, the danger of criterion-referencing is that one only plans for and assesses the observable, performatory and often superficial, trivial aspects of education and proceeds to record them in pages of tick boxes. Attempts to minimise these problems and to make criterion-referenced tasks manageable can easily result in generalised and inaccurate data that are meaningless.

Further, if aggregation is to occur, one has to be clear what exactly is being aggregated. For example one could be aggregating components that were so dissimilar to each other, in terms of levels of difficulty, as to yield a meaningless composite. This can be illustrated by an example in English. Let us say that I receive a Level 3 for spelling and a Level 5 for reading; if I aggregate the scores to give me an average level I will be awarded a Level 4. But what does this Level 4 really mean; what does it indicate? It is a meaningless number that tells me nothing about my specific achievements in the contributing criteria; I am unable to use the result for subsequent planning – the formative potential of criterion referencing has been lost. What if there was no parity in actual difficulty between a Level 3 in spelling and a Level 3 in reading (which is likely to be the case)?

The task, then, for users of criterion-referenced assessments is to select limited *sample* criteria that fairly represent a wider field, balancing specificity and generality whilst adhering to notions of: (i) *item discriminability* – the potential of the item concerned to be answered correctly by those students who have a lot of the particular quality that the item is designed to measure and the potential of the same item to be answered incorrectly by those students who have less of the particular quality that the same item is designed to measure; (ii) *item difficulty* – where the item is not so difficult that nobody answers correctly and not so easy that everyone answers it correctly.

## References

Gipps, C. (1994) *Beyond Testing*. RoutledgeFalmer, London.

Morrison, K. R. B. and Ridley, K. (1988) *Curriculum Planning and the Primary School*. London: Paul Chapman Publishing.

Morrison, K. R. B. (1993) Building progression into modular higher degrees in education. *British Journal of In-Service Education*, 19 (3), 5–11.