ROSBA’s Family Connections

Discussion Paper 1

Abstract: The Radford scheme belonged to a family of procedures known technically as “norm-referenced” assessment. The current system, called ROSBA, focuses on criteria and standards and belongs to the “criterion-referenced” family. In this Paper, something of the similarities and differences between these two families are outlined. It is also shown how ROSBA differs from the criterion-referenced testing movement in the U.S.A.

Author: Royce Sadler, Assessment Unit, January 1986

Note: This Discussion Paper has been produced by a member of the Assessment Unit of the Queensland Board of Senior Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. Reactions to and comments on its contents would be helpful.

© This material is copyright. It may be copied freely for the use of schools and teachers in Queensland. It may not be reproduced for sale without express permission.

ROSBA belongs to a family of procedures known technically as “criterion-referenced assessment”. The Radford scheme, on the other hand, belonged to the “norm-referenced” family. Each family has a long and honourable history, each family can do some things better than the other. This Discussion Paper sketches in the two families as background to what ROSBA is about.

It is well known that a raw score in isolation possesses little intrinsic meaning. For instance, to say that a student has received 73% in a test does not tell us whether the level of performance is good or bad. Among other things, the test may have been exceptionally easy, or the teacher prone to award high marks for mediocre work. In some subjects, 73% would constitute failure! Similarly, to say is some context or framework that allows an index of achievement to be interpreted.

Basically, scores are given meaning by providing some basis for comparison. The three common approaches to giving scores meaning are distinguished by their comparison points. Specifically, the three approaches involve comparing a particular achievement (a) with the achievements of other students who have attempted the same tasks (“norm-referencing”), (b) with explicit descriptions of the actual things mastered (“criterion-referencing”). The last one is important in formative assessment, and in the education of the intellectually and physically handicapped. Where improvement is slow or achieved only with difficulty, any move towards success, no matter how small, is attended with celebration. Attention in this paper, however, is confined to the first two approaches because they are the ones more applicable to mainstream schooling. Note that norm-referencing and criterion-referencing represent two different approaches to interpreting and reporting student achievement, not to two different types of test items. One cannot tell just by looking at, say, a test whether it is intended to provide norm-referenced or criterion-referenced data.

Norm-referenced assessment

Norm-referencing is the traditional method of interpreting scores. Some current (and most older) school reports record the student’s score, the class mean, and the place in class. More formally, scores may be
rescaled mathematically, and converted into new or “derived” scores that make allowances for such things as the difficulty of a test, and permit more uniform interpretations. The Tertiary Entrance Score in Queensland, under both Radford and ROSBA, is an example of a norm-referenced measure. It reports the performance of a student relative to that of the total age group. Any student with a T.E. Score of, say, 975 is judged to be better academically than 97.5% of the age group, and is therefore in the top 2.5%. Technically, a T.E. Score is a (slightly modified) *centile rank*.

There are two main attractions of norm-referenced assessment. First, it provides a single score that gives a concise summary of a student’s level of achievement, even when there is no detailed analysis or inventory of particular accomplishments. Second, it simplifies decision making in situations where selections on the basis of merit must be made. This is because norm-referencing greatly facilitates ranking. But norm-referencing also has some limitations. One is that in forming aggregates of marks, the details of specific competencies are not explicitly identified at all and marks are treated as a kind of currency. A second disadvantage is that when the common practice of grading on the curve is used, the proportion of students receiving each grade is fixed, and intense competition among students often develops. Finally, an exclusive focus on norm-referenced assessment makes it difficult to monitor shifts in the performance levels of successive cohorts of students unless the test themselves remain relatively fixed.

**Criterion-referenced assessment**

Criterion-referencing is an omnibus term that covers, broadly speaking, any attempt to interpret a students’s performance by referring, not to the performance of other students, but to specified domains of knowledge or behaviour. In principle, it should be possible for someone who knows a particular student’s score or achievement level to consult a catalogue that describes exactly what that student can do. Although the term “criterion-referencing” was not used until the early 1960’s, the basic idea is an old one, and motivated many widely used achievement scales in spelling, composition, drawing, handwriting, and vocabulary, dating from the turn of this century.

Part of the attractiveness of criterion-referenced assessment is the hope it holds out for a system of non-competitive assessment, in which students pit themselves against defined levels of achievement (which incorporate *standards*) rather than against one another. In doing so, of course, criterion-referenced assessment will inevitably discriminate among students, given an appropriate set of standards, and students of mixed abilities. But discrimination is not a main aim. In addition, having the nature of student achievement particularized has considerable utilitarian appeal. However in some subject areas, criterion-referencing turns out to be fairly difficult to put into practice, certainly more difficult than norm-referencing. It is therefore not surprising to find great variety in the research and development thrusts trying to find solutions to the problems. This is not the place to survey this variety, but it should be observed that the ROSBA philosophy (a) is located firmly within the criterion-referencing tradition, but (b) is sufficiently distinct from the most fully developed existing varieties in the U.S.A. for it to require independent developmental work.

Norm-referenced and criterion-referenced assessment are often portrayed as if they were mutually exclusive policies. Conceptually, it is easy (and neat) to separate them but in practice, they cannot be entirely divorced from each other, as we shall now see. Norm-referencing, in its ideal and abstract form, takes no account of the quality of student performance in an absolute sense because it is primarily concerned with orderings among students, that is, with determining which students are better than others, and by how much. Theoretically, with purely norm-referenced assessment, the average achievement level could fall year by year to some abysmally low level and no one would be any the wiser. In practice, parents and employers become concerned when students leave school after 10 or 12 years and are unable to read or write a letter, or have little appreciation of science, the culture, or the environment. This acts as a weak and indirect criterion-referenced check on what schools help students achieve. A more potent force is, of course, the integrity of teachers and the requirements set out in syllabus documents.
Similarly, a criterion-referenced system would soon come unstuck if the standards were set so absurdly low that all students were classified as “excellent”, or if standards were set so high that no one ever achieved them. Educators do have notions of what can reasonably be attained by students. Standards specifications that are designed by subject matter experts should reflect the range of achievements expected at a particular level of schooling. However, the specifications must not nominate in advance the proportions of students to be given the different grades. In practical terms therefore, standard setters draw on what they know students can achieve (the “norms”). Once standards are defined, the norms become irrelevant.

Criteria and standards distinguished

As terms, standards and criteria are often used interchangeably, both in ordinary conversation and in discussions about assessment. However, a distinction can be made. Not only does it have some (but not universal) backing in the educational literature, it also turns out to be a very useful one in that it breaks the process of teacher judgment into two stages. First, the criteria have to be identified, then standards on the various criteria specified.

A criterion (plural: criteria) is a property, dimension, or characteristic by which something is judged or appraised. For example, originality, neatness, and accuracy are three criteria that could be used in assessing student projects. A standard is a fixed reference point for use in assessing or describing the quality of something, or as a goal to aim for. The distinction between a criterion and a standard can be clarified by means of an example, in this case one from outside education. Suppose that in testing a bicycle helmet for safety the authorities stipulate that its impact resistance must be at least 75 units for the helmet to be given A-Grade rating. Impact resistance is the criterion; 75 units is the standard, or the minimum level to be satisfied for an A-Grade rating. A helmet that tests, say, 80 units obviously meets the standard for impact resistance. Other criteria might also be important, such as weight, and visibility. Each of these criteria will have associated standards for an A-Grade helmet. B-Grade helmets would use the same criteria, but the standards would be lower. The combination of the standards on the three criteria might be called collectively “the overall standard”. Under ROSBA, the combination is called the “exit Level of Achievement”.

In their purest form, standards are descriptions or other specifications of performance levels that are free from any references to the performance of the “typical” student, the proportion of students expected to achieve a given level, or the particular age or stage of schooling at which a certain level of performance is thought to be reasonable. An example of such an “absolute” standard comes from typing, where the main criteria are speed and accuracy: “The student can type at 70 words a minute with 95% accuracy”. (The difficulty of the prose, the definition of a word, and how mistakes are counted all conform to accepted conditions of testing, and are stated in the syllabus). Although there are reasons for believing that absolute standards can be devised in all areas of the curriculum, in most subjects their proper formulation will require ingenuity and persistence. This is one reason that they should probably be developed under the auspices of Subject Advisory Committees and incorporated into syllabuses. (Because they are generally difficult to construct, the temptation is often to avoid them and use norm-referencing instead.) Measurement may or may not be necessary, the essential point being that teachers will be making judgments about qualities and quality. Sometimes (as in the typing example) counting or measurement will form an essential step in determining quality or competence. In many other areas, too great a preoccupation with numbers and scores may get in the way of determinations of quality. The problem of finding workable ways to fix, define, and promulgate standards is by no means trivial, and its successful solution is likely to involve members of the Assessment Unit, Inservice Team, Review Office, and Subject Advisory Committees in extended dialogue and experimentation.

The definitions given above show that the primary focus of ROSBA is on standards rather than on criteria. However because standards presuppose criteria, an assessment system based on standards incorporates necessarily the concept of criterion-referencing. For this reason, the new scheme could be referred to generically as “standards-based assessment”.
It is necessary at this point to make crystal clear just what types of criteria form the foundation of a criterion-referenced assessment system. Without that, communication among different parties involved in education will be at cross purposes, and misunderstandings will multiply.

Criteria are the levers by which judgements are made and defended. Such judgments include classification of things or people, determinations of the most appropriate courses of action, and appraisals of quality or achievement. It is quite obvious that all of the decisions required for running and educational system employ criteria. Indeed, it is said (quite correctly) that using criteria is not particularly novel, because teachers have always had criteria in mind when setting tests. In the normal course of events, teachers have to decide which assessment instruments to use (using the criteria of objectivity, efficiency, relevance, and the like) and which items to include in a test (using such criteria as content validity, specified, difficulty level, and clarity). Other criteria are used by persons involved in the review process for judging the adequacy of Work Programs. But criterion-referenced assessment is concerned exclusively with the criteria that can be used for evaluating student achievements or competencies as outcomes, and presupposes good Work Programs and good assessment.

Standards-based assessment and Radford

ROSBA has been variously described as a “fine tuning of” or a “radical replacement for” the Radford scheme. Both positions have some validity. ROSBA retains and strengthens many of the Radford initiatives, especially the shift to school-based curriculum development, teacher assessment of students, and a distributed responsibility for making checks and balances to the system as a whole. However, the most radical element in the new system is a shift away from norm-referenced assessment, and towards criterion-referenced principles.

One of the motivations for this change of direction (it represents more than a shift in priorities) was a determination to reduce or eliminate some of the undesirable effects that are associated with fierce competition. This change in direction will be ultimately of greater significance than the mechanics of the changeover. At present, teachers, schools, the inservice team, and review officers are more-or-less submerged in a deluge of paper, meetings, work programs, criteria and standards, assessment instruments, student folios, and some confused signals emanating from a number of sources as to what the whole exercise is about. (In addition, the requirement of tertiary institutions for selection mechanisms engages teachers in an exercise that is clearly more norm-referenced than criterion-referenced, and the duality is causing some tensions.) When the fold subsides, there will be not only a clearer conception of what standards-based assessment is, but a considerable amount will be known about how to put it into operation.

Standards-based assessment and the criterion-referenced testing movement

Standards-based assessment is clearly a local variety of criterion-referenced assessment, and has drawn on the literature and experience elsewhere for its inspiration. In particular, it shares a concern for definitive grade specifications that set out what the various Levels of Achievement are to consist of, taking into account the various types of objectives and outcomes. But there are some points of differences. Apart from the organisational arrangements set up by the Board of Secondary School Studies to achieve comparability and consensus (while at the same time supporting school-based development), standards-based assessment is less concerned than the criterion-referenced testing movement generally with measurement as such, and with determining cut-off scores for mastery. It recognises the key role of the qualitative professional judgments of classroom teachers, both in the evaluation of individual pieces of student work, and in integrating that information to decide on a level of achievement. It puts great emphasis on the identification of criteria, the determination of standards, and the generation of policies that specify permissible trade-offs among different components. (These points are amplified in the two companion Discussion Papers “Defining Achievement Levels” and “A Standards Schema”).
This should not be taken to imply that teachers’ judgments are assumed to be, *without exception*, infallible, highly reliable in the psychometric sense, or even comparable from teacher to teacher. What it does assume is that teachers by virtue of the relationship to students are in the best position to make judgments about their students’ work, simply because they have access to the fullest information. But more work needs to be done in improving those judgments.

As we are finding out, the theory of a standards-based assessment is disarmingly simple but the practice is extraordinarily difficult. But having got this coveted ball into our court, we are going to see where we can hit it to maximize the good effects.
Author/s:
Royce, Sadler D

Title:
Discussion Paper 1. ROSBA's Family Connections

Date:
1986

Persistent Link:
http://hdl.handle.net/11343/115864