Kroniek

Research into monitoring* national standards of educational achievement

Deze tekst is geschreven door Dr. M. Tyerman, Head of Educational Research Section, Division for Educational Documentation and Research van de Raad van Europa, n.a.v. een 'European Contact Workshop', die onder auspiciën van de Raad van Europa van 13–18 juni 1976 in Berkshire, Engeland, is gehouden.

1. *The theme*

The purpose of the workshop was to help leading researchers in member states develop methods by which they might regularly assess (monitor) the achievement of pupils at a national level as a means of studying the processes and effectiveness of their school systems.

This paper is an attempt to summarise the proceedings in a non-technical way that does not refer to particular papers. A more detailed account together with the workshop documents edited by the organiser (Dr. R. Sumner) is expected shortly to be published by the National Foundation for Educational Research Publishing Company, Jennings Buildings, Windsor, Berkshire, England.

2. *Why monitor?*

The last twenty years have seen an enormous growth in educational opportunity. School provision and ancillary services have increased and there is growing emphasis upon education for adults and the concept of lifelong education. In some European countries the cost of such facilities amounts to 7 % and more of the gross national product. But with greater opportunity and increased expenditure there is mounting criticism of educational systems, disappointment with the results of increased schooling, and disenchantment with the view that a better education for all will necessarily lead to a better life for all.

In such a situation there are powerful utilitarian arguments for monitoring standards within schools. Taxpayers should be shown that their money is well spent, self-interest demands effective education, and national pride and international status is bound up with the educational level of the electorate. But, if taxpayers are not getting value for money, what then? Not all governments welcome weakness in their systems being exposed, and when defects and limitations are revealed it is much easier to publicise them than to remedy them.

Additionally, there is the need to discover the special educational problems of socially disadvantaged groups, such as poor indigenous workers and certain immigrants and migrants, so that their situation may be improved. The generally slow progress of their children in school has led to a re-examination of the concept of equality of educational opportunity, and to an awareness of a need to assess accurately and continuously the effects of change. This in its turn has furthered the shift of attention from 'inputs' into the educational system such as increasing the number of schools, lengthening schoollife, lowering pupil-teacher ratios, and devising new curricula, to 'outputs', to assessing the effects upon the pupils of such developments. Equal progress is now seen as the criterion of equal opportunity, not equal resources.

3. *What should be monitored?*

A country's educational system reflects its social and political philosophy. The standards to be monitored should, therefore, be those which indicate most clearly the progress that is being made towards achieving its goals. But these aims are usually described in vague philosophical language rather than in clearly defined behavioural terms that relate to limited specific objectives. Progress towards such objectives can usually be measured, that towards aims cannot.

One cannot assess in any valid way the progress that a national sample of pupils is making towards such aims as 'the fulfilment of potential'. In practice, therefore, researchers have tended to concentrate upon a limited number of cognitive skills, especially reading and arithme-

* 'To monitor' is te vertalen met 'evalueren'; 'monitoring standards' is ook als 'niveau-bewaking' te lezen.

tic. This has the advantage of relative simplicity and of focusing upon immediate concerns. For, whilst discussions of educational aims may very properly centre on developing responsible citizens of good character, the chief criticism made of schools is usually that many of the children leaving them cannot read or write properly. Furthermore, there are reliable and valid objective tests of reading and arithmetic which can be easily administered and marked.

But school subjects that are the easiest to measure are not necessarily the most important nor the most indicative of general attainment. For example, the ability to do everyday arithmetic or spell correctly could properly be regarded as a valid indication of whether or not a primary school system was effective, only if the school system considered such skills as being of first importance.

In one member country an official handbook encourages a deliberate move away from learning factual material towards a fostering of curiosity in the child and developing his capacity to discover things for himself whilst ensuring that the fundamental skills of reading, writing and arithmetic remain as basic elements of the school course. How do you define 'curiosity'? How do you define 'the capacity to discover things for himself'? How do you assess it?

It can be argued that the only fair way to judge the effects of schools is to assess the work that is being done in them. This would demand evaluative techniques that accord with the goals and curricula of the schools, but there may be little agreement in explicit terms on what those goals should be, whether in a centralised system, such as that of France, or one that is decentralised, such as that of England.

4. *How should standards be measured?*

The first step must be to decide which educational objectives are sought, and what their relative orders of importance might be. They should be defined preferably in ways that can lead to objective as well as subjective assessment. Such measurement should not be limited to those concepts that are deliberately taught but should encompass the various facets of school life and the incidental learning from the 'hidden curriculum'. Questions should then be designed or selected that relate to these objectives and whose answers would show which items of knowledge have been assimilated and which skills have been learned. There should be a large enough number of questions to prevent chance affecting the score and they should be so constructed that they could be fairly regarded as a random sample of all the questions of that type. In other words they should be generalised and the error in the generalisability of the questions must be estimated before conclusions are drawn.

The usual technique has been to apply standardised objective norm-reference tests over a range of subject areas. These tests generally contain a large number of questions (fixed content) that have been tried out (standardised) on a representative sample of pupils of particular ages. Such standardisation enables the test constructor to retain only the most disciminating questions, and to calculate average levels of attainment (norms) for pupils of different ages. The use of such tests allows national or local standards to be assessed and comparisons between different areas of the same country to be made.

Norm-reference tests have, however, intrinsic limitations and are unsuitable for judging changes in pupils' attainments over a period of time. They measure a very random range of skills and they cannot take into account the variety of curricula and approaches followed in different schools or by the same schools in different years. In other words they can lack content validity.

Most studies of achievement have tended to be cross-sectional rather than longitudinal, for example, they have assessed a sample of 8-year-old pupils and a sample of 10-year-old pupils. The results are taken to show the attainment of pupils at 8, and at 10, and the differences between the standards of the two samples to show the work of the two years between 8 and 10. Leaving aside such questions as whether the samples of children studied are truly representative of all pupils of that age, and whether the tests being used give consistent results and measure what they are supposed to measure, such an approach does not allow cumulative processes to show.

In a longitudinal study the progress of the same children is assessed at intervals of time using the same tests or different tests. If the same test is used its content may be outdated after the first occasion and there is a practice effect to be calculated. If a different test is employed there is the problem of ensuring comparability between it and the earlier test. This has led to the development of item banks.

If the same questions cannot be used again and again a large number of questions or items of comparable difficulty is required. Such a collection of items with calibrated data on their measurement characteristics constitutes an item bank. Two tests made up of separate items from the bank can be interpreted one in terms of the other. Year by year test items are added to the bank or dropped when their usefulness has been outlived. 'Multiple matrix sampling indicates that such a procedure is valid and that a test so made has no major disadvantages compared to a test with a fixed content. Furthermore, such item banks are particularly well adapted for criterion reference evaluation.

A test is criterion referenced when provision is made for translating the score into a statement about the behaviour to be expected from a person gaining that result. A norm-reference test indicates how a person or group compares with another person or group. The criterion reference test describes what he or they are able to do either by describing the actual level of performance or by an expectancy statement predicting performance in a situation unlike that of the test, for example, interpreting the results of a reading test by newspapers that a child with such a score might be expected to understand.

It is increasingly agreed that the results from normative tests should be supplemented by data from criterion reference tests, and that these details should be balanced by information that is based on first-hand observation. To use the current phraseology, evaluation based on the results of tests must be illuminated by a more subjective

yet systematic observation of pupils' work over a wide range of subjects, and by a judgment of the qualitative aspects of pupils' own work. An appraisal of attainment in school subjects is needed at a deeper and wider level than the marking of responses to tests, and some judgment of progress in personal development is essential, for example, in such qualities as emotional maturity, sociability and perseverance.

5. *Implications for governmental policy*

Two problems in particular face an administrator. How to ensure that the best use is made of limited financial resources, and secondly, having made the provision, how to discover whether the additional resources are in fact making any difference.

There is a notable difference between the ways in which decisions on education are reached and those in certain other areas. Usually economists try and compare the advantages of increasing or directing expenditure in different ways, but in education it is usually difficult to foresee exactly what will result from any change in financial or other provision. Attempts to measure academic progress and personality development and then to relate it to the type and extent of the education received have given different results.

There is little general agreement on what are the conditions which contribute most significantly to the knowledge or skills of pupils. Most of the situations (variables) that have been examined in such investigations tend to be closely associated (correlated) with each other. To try and assess their separate and collective weights highly sophisticated statistical techniques have been employed. In general, such analysis has suggested that home background may be more potent than conditions within the school in determining the attainment of pupils. Thus achievement may be more a reflection of pupils' family and social circumstances than of the education given them.

In a country with an educational system where there is central control of curriculum content as well as of resource availabilitiy, the development of valid performance measures might well make central decision-making more effective. In countries with an overtly federal organisation of education performance monitoring would probably be more significant at provincial than at national level. In countries where the organisation of education is diffused the results of monitoring seem likely to appear in different ways. For example, any national assessment model must be able to respond, at least in part, to local as well as to national requirement and uses. However, in six recent situations in which evidence of pupils' attainments influenced governement action regarding education, it was clear that in no case was their performance the sole determining factor.

Changes in educational policy and practice require changes in methods of assessment; similarly new techniques in evaluation might make it possible to specify the objectives of particular policies more precisely. This whole issue of the relation between monitoring and governmental policy raises philosophical and ethical problems. Is monitoring concerned with the goals of education or with the goals of an educational system? Is there a danger that national monitoring might become political involvement? And, are there situations which would warrant researchers withholding the results of their enquiries from the public? Is it justifiable to carry out national surveys purely for use within ministries? And if not, how can the information obtained be presented in a way that is useful to teachers and other members of the public? If policy should be determined by the electorate, how can the results of national surveys be presented in a way that is meaningful to them?

6. *Results and follow-up: comment by the Secretariat*

The discussions of the workshop indicate that whilst there is widespread and growing concern over educational standards, there is at the same time a lack of agreement on the relative importance of different educational objectives and on what constitutes an acceptable line of progress towards them. In such a situation, sweeping statements about the success or failure of schooling in general and of particular practices within educational systems do not seem justified.

There are plenty of opinions about educational standards but few facts. Monitoring attempts to supply those facts.

Monitoring is here to stay. At Windsor there were reports on national surveys from France, the Federal Republic of Germany, Ireland, Switzerland and the United Kingdom, and similar enquiries have been undertaken or are being planned in other countries. A knowledge of the basic principles of monitoring and how it can be carried out at national or local level if essential for administrators, inspectors and teachers.

It is now true to say that:

1. Many of the technical problems that are faced in designing national surveys and measuring pupils' achievement have been solved. The Rasch model, Bayesian theory and generalisability theory are particularly important here.
2. Techniques of testing are now so sophisticated that by using banks of test items, children in one age group can be compared with those in another even when they do not answer the same questions.

Yet these developments are only the first steps in an exceedingly difficult and complex research process. It is still not known how to measure accurately

1. pupils' original work, nor
2. their personality development,

though the growth of creativity and emotional maturity are fundamental educational goals. And whilst it is clear that monitoring makes great demands upon money, time and expertise, there are no figures available to give precise details.

In monitoring the following two points are also evident.

1. Scientific and technical aspects cannot be divorced from educational and socio-political issues. Such matters as, for example, confidentiality must be resolved in the early planning stages and the co-operation of teachers obtained
2. The reasons for undertaking the enquiry must be clear to all concerned and the objectives of the schools or system being investigated must be understood. These considerations will determine the ways in which the information is to be collected, and how it is to be analysed and interpreted. The statistical criteria must always be balanced by educational judgments. For this reason methods adopted and conclusions reached in one country may not be applicable in another country without considerable modification.

Advances in monitoring will depend upon field work in which data is analysed in different ways and surveys are repeated with different samples of pupils. A variety of approaches will be required. For this work the skills of researchers are not enough. They need the co-operation, special knowledge and active help of ministry officials, administrators, inspectors and teachers. The extent to which this is given will depend in its turn upon how far monitoring is seen as useful by such people in determining both the day-to-day problems and the long-term policies with which they are involved.

To enable educators from various countries to discuss this point and the other issues raised in this paper at a practical level a meeting arranged by the Council of Europe might be helpful. Readers who would like to suggest that such a meeting be held are invited to inform the Secretariat (Division for General and Technical Education).