

Udo Bude & Keith Lewin (eds.)

Improving Test Design

Vol. 1 – Constructing Test Instruments, Analysing Results and Improving Assessment Quality in Primary Schools in Africa



Cover photo: Udo Bude/DSE

Impressum

published by:

Deutsche Stiftung für internationale Entwicklung (DSE)
German Foundation for International Development
Education, Science and Documentation Centre
Hans-Böckler-Str. 5
D-53225 Bonn

editors:

Udo Bude/Keith Lewin

editorial work & typesetting:

txt redaktionsbüro ebert & schmid, Lünen

printed by:

Druckerei Brandt GmbH, Bonn/1997

DOK 1788 a

ISBN 3-931227-27-8 (complete edition)

ISBN 3-931227-28-6 (Volume 1)

1. Introduction	5
2. Theoretical and Practical Aspects of Item Construction for Primary Science Subjects...	10
2.1. Criterion-Referenced Assessment - Panacea or Palliative?	10
2.2. Test Development - Designing Tests and Presenting Results.....	24
2.3. Developing and Administering Tests in Primary Science and Agriculture - A Practical Exercise.....	38
2.3.1. The Item Writing Task.....	41
OVERVIEW.....	41
Info 1: Primary Education Syllabus Kenya - May 1992 (abstracts).....	42
Assignment 1: Basic Instructions	50
Info 2: Advantages and Disadvantages of Multiple-Choice Questions.....	51
Info 3: Format of Multiple-Choice Items	52
Info 4: Suggestions to Consider in Writing Higher-Order Items.....	52
Assignment 2: Item Writing and Gender Aspects	53
Info 5: Gender Stereotypes in Science Teaching/Learning	53
Assignment 3: Individual Review of Test Items and Prediction of Outcomes.....	54
Assignment 4: Further Review of the Draft Test	54
TRIAL TEST PAPER: Science and Agriculture (incl. Making Scheme).....	55
Assignment 5: Item Prediction	69
Assignment 6: Administration of Test	69
Info 6: Summary of Item Analysis Procedures.....	69
Assignment 7: Preliminary Analysis of Data	72
Info 7: A Simple Measure of Reliability	73
WORKING GROUP RESULTS.....	73
Outline for Interview with Science/Agriculture Teachers	73
Reports on Interviews in Primary School	76
Reports on Test Administration.....	80
Reports on Test Results in Schools.....	81
2.3.2. Reviewing the Item Writing and Testing Process.....	85
Deciding a Specification for the Range of Items to be Constructed	85
Development of Test Items	86
Trial of Test Items in Schools.....	86
Marking of Responses	86
Analysis of Item Characteristics.....	87
Concluding Remarks.....	88
2.4. Bibliography.....	90
3. Primary School Examinations in Kenya with Special Reference to Item Construction for Science and Agriculture	92
3.1. The Use of Examination Results for Monitoring Performance of Schools, Districts and Provinces.....	92
3.2. Testing and Monitoring Procedures Developed for Primary Schools	96
3.3. Developing Tests for Complex Cognitive Processes	105
3.4. Skills in the Construction of Science Tests	110
3.5. Technical Efficiency in Constructing Test Items.....	117
4. APPENDICES.....	124
4.1. Examinations Policies to Strengthen Primary Schooling in African Countries.....	124
4.2. Criterion-Referenced Testing. Rationale for Implementation (Abstracts)	139
4.3. Evaluation of Examination Needs of Primary & Secondary Schools in Namibia (Abstracts).....	146
4.4. Primary Science School Leaving Examination (Abstracts)	151
4.5. Learning Competencies For All. Essential and Desirable Learning Competencies for Standards 4, 5 and 6 (Abstracts)	155
4.6. Science in the National Curriculum - Attainment target 9: Earth and Atmosphere.	162
4.7. Boat Building (The Properties, Classification and Structure of Materials) - Worksheet.....	163

1. Introduction

Udo Bade & Keith Lewin

As a follow-up to the World Conference on Education for All in 1990 several countries in southern and eastern Africa started analysing and revising their examination systems with the ultimate aim of improving the quality of teaching/learning. Examinations in countries of the region are a well established feature of the educational systems reaching back to colonial days. The achievements of pupils during the primary school cycle are in most cases tested in end-of-cycle examinations, whereby the performance of a pupil is tested in comparison with other pupils sitting the examination (norm-referenced tests). Such assessment procedures do not provide the full picture and fail to deliver sufficient information about the success of an education system in imparting those skills and competencies as laid down in the curricula.¹ In order to find out how proficient a pupil is in a particular subject, concept or skill without referring to other norm groups of pupils a different approach is required. This may have several elements which include greater use of school based assessment, evaluation of project work and techniques of continuous assessment. It may also seek to define competencies in terms of criterion statements against which performance can be judged. Criterion-referenced assessment is attractive since it sets standards that do not depend on the performance of other pupils and should provide reliable data on what has been achieved during the primary school years. Some countries in the region e.g. Botswana and Swaziland are already experimenting with this kind of approach to assessment.

¹ See Kellaghan, Th. & Greaney, V. (1992) Using Examinations to Improve Education. A Study in Fourteen African Countries. World Bank Technical Paper No. 165. The World Bank, Washington D.C.

Applying more comprehensive concepts of assessment is a significant step towards the improvement of the quality of teaching and learning. National examination systems are mostly designed to judge the individual pupil's achievements for selection purposes and to deliver comparative information about the performance of individual schools and regions. They are often an unreliable guide to actual levels of achievement.

Three aspects have to be taken into consideration when constructing most forms of assessment instruments:

- (1) the validity of assessment instruments (do they measure what they claim to be measuring? do they predict future performance adequately if they are to be used for selection?);
- (2) their reliability (do they work a consistent measure of performance which could be repeated with similar results? are measurement errors reduced to acceptable levels?);
- (3) their technical efficiency (is the system secure, cost-effective, or as appropriate time scale, free of bias towards or against different groups of candidates?).

For most pupils in the region primary education ends after seven to eight years with a national examination. The more selective such examinations are the greater is the attention and importance given by society, because the results of such annual exercises determine significantly the future of many children and the hopes and ambitions of many parents for their offspring.² The outcome of the examinations also have severe repercussions on the schools on learning and teaching methods, and on the teachers' role in local communities and within the education system (see box: »Vihiga plan« is the way out/Fortunes change for Kikuyu).

² See Dore, R. P. (1976) The Diploma Disease. Unwin Education. London.

EAST AFRICAN

Standard

Established 1902

COMMENT

‘Vihiga plan’ is the way out

PARENTS in Sabatia Division of Vihiga District have embarked on an ambitious programme to improve education standards in the area.

Part of the master-plan lies in setting up an in-service training centre for secondary and primary school teachers and their deputies. Those who attended the leaders' meeting on how to improve education in the area made proposals of providing lunch to all examination classes in future.

However, whereas there is no so much novelty in advocating for managerial skills for headteachers, or even starting schools' feeding programme, it is important that parents are ready to initiate the scheme on a harambee basis.

Going by the results of the last year's Kenya Certificate of Primary Education (KCPE), one can just understand the urgency of having efforts to improve education in Vihiga. Out of 56 districts and municipalities, Vihiga was number 41.

By all standards, those results were not rosy, but perhaps the most important aspect is that parents are prepared to reverse the situation.

But even then, they are throwing a big challenge to the Ministry of Education and the Teachers Service Commission (TSC) whose responsibility is train and appoint of headteachers.

In the past, appointment of headteachers in both primary and secondary schools has been done in total disregard of experience and competence of headteachers and their deputies. Poor supervision of schools added to the problem.

As the leaders' meeting at Vokoli Girls Secondary School noted, there is paucity in supervision of headteachers, who more often become their own masters and oppressors of parents and other people who refuse to toe the line, or who question their decisions.

It is also important for the Ministry of Education to acknowledge community based programmes, and support them.

Fortunes change for Kikuyu

By STEPHEN MUMBU

The performance of the Kenya Certificate of Education has greatly improved in Kikuyu Division, according to the local education officer, Alan Mwangi.

He said the division which used to come last out of seven in the district was now number three.

Kikuyu came third after Lari and Githunguri in last year's KCPE results.

Addressing a prize-giving ceremony at Kikuyu Township Primary School recently, the officer teachers, pupils and parents to work harder to get better results.

The remarking of Mzee Giteu Primary School KCPE papers elevated it up from 110th to fourth position with a mean score of 450 marks in the district's 200 schools.

The Zonal Inspector of Schools, Mr James Wamwari and the Teachers Advisory Centre (TAC) tutor, Mrs Kingdane Kiamuthi, praised teachers for their work which had enabled Mungu to produce six of the top 10 schools and also six of the top 10 pupils in the district.

Kabeta produced the best candidate in Samuel Muryu from Mbia-ini Primary School.

Daily Nation
4/3/95

Tuesday, February 28, 1995

Despite the great attention paid to matters of testing and examining at the end of primary education, the examination results rarely provide sufficient information on the effectiveness of the educational system to make confident judgements on educational quality and learning problems. If we are interested in influencing the teaching-learning process in the classroom positively we may have to start with improvements in the ways pupils are assessed. This can indicate what is not being understood and assimilated, and can point the way to strategies to improve levels of achievement.

Any assessment system provides opportunities for teachers, pupils, and educators (e.g. curriculum developers, examination/testing specialists) to monitor progress and learn from failures as well as successes. However, if large parts of examinations consist mainly of straightforward recall questions where pupils are simply asked to reproduce from memory, opportunities for a more comprehensive assessment of learning outcomes are lost and much teaching will follow objectives narrowly defined by a restricted range of questions. The mere recollection of facts or names does not give any hint of the pupil's problem-solving skills, often so strongly emphasized in the primary school curricula as essential outcomes of learning.

Repetition of whole sections from textbooks fails to indicate whether learners are able to apply their knowledge and skills to different situations. Poorly constructed multiple-choice tests leave much space for guess work and may encourage rote learning. They need to be replaced by a broader concept of assessment testing more complex cognitive processes. Consequently, the first step in a reform of conventional examinations is to analyse existing tests in terms of whether and to what extent they contain items requiring higher-order thinking and application. The next step is to develop or improve test items which examine pupils' abilities to apply what they have learned to less familiar situations and problems or which require them to link events or facts to each other in a consistent way.

The development and design of test items is mainly the domain of examination specialists and/or curriculum developers. The work of all those involved in examinations should, but often does not, include a strong research component in order to find out what kind of assessment procedures deliver the most valid and reliable results efficiently and have the greatest positive influences on classroom teaching. Educationists responsible for the design and analysis of national examinations in eastern and southern Africa are trying hard to improve their assessment systems and to answer the challenges of new ideas and requirements.³ During recent years new subjects were added to previous core subjects, i.e. like Languages and Mathematics. In many primary school systems History, Geography and Civics or Social Studies have been included, along with science based subjects like Science, Agriculture, Environmental Science, Home Economics. Most or all now feature in the national examinations. Writing meaningful tests for the more practical subjects which go beyond recalling facts requires familiarity and experience with the respective school curriculum as well as with the possibilities and limitations of different types of assessment. Often the demanding curriculum objectives of subjects like Science or Agriculture are not easily tested in an appropriate way. Developing test items assessing pupils' understanding and application of knowledge and skills to new situations require sophistication in constructing assessment instruments even where testing is restricted to multiple-choice questions for reasons of cost and administrative feasibility.

³ See Njabili, Agnes F. (1993) *Public Examination: A Tool for Curriculum Evaluation*. Mture Publishers, Dar-es-Salaam.

Education experts responsible for revising national examinations to improve the quality of education can learn from the approaches adopted to reform in different countries in the region, particularly in the following areas:

- widening the range of skills tested and the instruments used;
- redesigning examination items to include more which test skills of higher-order thinking;
- gradually shifting the basis of testing from a norm-referenced to a more criterion-referenced system (measuring pupils' success or failure in relation to criteria which represent competencies independent of the performance of other pupils);
- considering the possibilities for introducing continuous assessment alongside or instead of terminal examinations, and
- using pupils' records and profiles which can reflect the acquisition of demonstrated competencies of a wider range of different types than can conventional examinations.

One country in the eastern and southern African region has over many years spearheaded new developments in using national examinations for monitoring and meaningful assessment purposes. Kenya started reforming primary education examination in 1974 with the declared aims of making the examination more relevant, improving the quality of education and ensuring greater equity in the access to secondary schools. Two major strategies for reform were employed; changing the content of the examination papers, and introducing an information-feedback system.»It was hoped that changes in the questions set would make the CPE more relevant as a leaving examination, more equitable to pupils in less-privileged socioeconomic groups, and more reliable as a selection instrument. The introduction of an

information-feedback system would, it was hoped, do something to improve the overall quality of the primary school system and to reduce quality differences between high performing and low performing schools.«⁴

⁴ Somerset, Anthony (1988) Examinations as an Instrument to Improve Pedagogy. In: Heyneman, Stephen P. & Fägerlind, Ingemar (eds.) University Examinations and Standardized Testing - Principles, Experiences and Policy Options. World Bank Technical Paper No. 78, Washington DC, pp. 171-194, p. 174.

The experience and expertise developed in Kenya over nearly twenty years is therefore very relevant to the present discussion on revising national examination systems in other countries of the region. The Kenya Certificate of Primary Education can serve as an example of possibilities and as entry point to become acquainted with assessment techniques at the end of primary school in the subjects SCIENCE and AGRICULTURE, and for monitoring primary schools in general.

The success of the reform of the primary school examination system has not been achieved without problems. Over the years Kenya has become more and more »exam-ridden«. The results of primary or secondary school examinations receive more and more public attention. Even to the extent that candidates' results are announced over public radio! Regions and school districts compete vigorously to top the lists in the national examinations, very often to the disadvantage of pupils who are unlikely to score highly and those who succeed but do so as a result of long hours of repetitive 'cramming'. The incentives to cheat and find illegal means to pass the test have also created many problems. Heavy emphasis on examinations leads to a neglect of the broader pedagogical tasks of the schools. The school curriculum may be only taught according to the importance of examination subjects and other aspects of the national examination ignored. The »examination tail is wagging the curriculum dog«! John Keeves reminds us of the real purpose of national examinations, »There is little doubt that a national examination has a substantial influence on the teaching that occurs in schools not only during the year at the end of which the examination is held, but in all years that have gone before... it is important that the examinations should have beneficial effects on the teaching and learning that takes place at all earlier stages of schooling«.⁵

⁵ Keeves, John P. (1994) National examinations: design, procedures and reporting. UNESCO: IIEP, Paris, p. 98.

The design and conduct of national examinations is therefore not an affair of one group of specialists alone. Examination specialists, curriculum developers, psychometrists and teachers have to cooperate to maximize the beneficial influence of the examinations on practical teaching and avoid excessive testing and examination preparations in the schools. Despite all good intentions on behalf of those setting the national examinations it seems to be extremely difficult to prevent examination »fever« once such examinations have gained too much importance in society as the means through which credentials are obtained, promotion through the school system rationed, and jobs in the labour market allocated.

The manual on IMPROVING TEST DESIGN tries to assist those educationists who are responsible for the design, conduct and control of national examinations in developing higher quality assessment instruments which can provide better information on pupil achievement, a fairer basis for selection, and influence learning and teaching to improve educational quality. It can also be used as a kind of source for training those assisting in the development or processing of examinations. The manual consists of two parts:

Volume 1: Constructing Test Instruments, Analysing Results and Improving Assessment Quality in Primary Schools in Africa,

Volume 2: Assessment of Science and Agriculture in Primary Schools in Africa; 12 Country Cases Reviewed.

Both volumes are the result of two pilot training workshops in Kenya dealing with the development of test items for Science, Agriculture, and Environmental Science and the use of national examinations for improving the quality of primary education. These workshops were

jointly organised and conducted by the German Foundation for International Development (DSE), Education, Science and Documentation Centre, and the Kenya Institute of Education (KIE) in cooperation with the Kenya National Examinations Council. Participants came from different countries in eastern and southern Africa. Each country invited was asked to nominate one curriculum developer and one examination specialist, thus guaranteeing that the curriculum aspects for Science, Agriculture, Environmental Science were equally considered with the examination requirements.

Volume one deals with the practical aspects of test construction, analysis and the improvement of assessment procedures. In addition Prof. Keith Lewin (University of Sussex) takes up some of the theoretical aspects, especially the possibilities and limitations of criterion-referenced assessment and test development in general. The practical exercise on developing and administering tests draws to a large extent on the experiences with assignments carried out during the second training workshop in Kenya, where participants designed test items, conducted tests in primary schools and analysed the test results in groups. The experiences of twenty years of examination reform are presented in two analyses authored by education specialists from the Kenya National Examinations Council and the Kenya Institute of Education. Finally, abstracts and papers are documented indicating the discussion and direction of examinations and test constructions in eastern and southern Africa.

Volume two starts with an account and analysis of the present situation regarding primary school leaving examinations of countries in the region. Detailed information for each country is provided in a tabulated overview illustrated by original examination papers in Science, Agriculture and Environmental Science mainly from 1993 and 1994. In few countries Science and Agriculture feature only as part of a larger »General Paper«. In these cases the items covering Science or Agriculture have been included in the documentation. South Africa and Namibia are also included, although presently they have not yet started end-of-primary examinations, but discussions on the composition and design of examinations are in progress. Furthermore, examples of follow-up communications after the examinations from different countries are presented.

2. Theoretical and Practical Aspects of Item Construction for Primary Science Subjects

2.1. Criterion-Referenced Assessment - Panacea or Palliative?

Keith Lewin

Assessment stands at the heart of effective school systems. Without an adequate system of assessment selection will be difficult to legitimate, certification will carry a wide range of meanings, monitoring of school performance will be difficult, and diagnosis and remediation of learning problems will be haphazard. Approaches to assessment vary. This chapter addresses some issues concerned with the long running debate about the relative virtues of norm- and criterion-referenced approaches to assessment and examination.

Many developing countries inherited public examination systems that were essentially norm-referenced. Since the primary purpose of the majority of these examinations was and is the selection of pupils to higher levels of education this is not surprising. Much recent thinking has tended to stress the value of criterion-referencing approaches to both public- and school-based assessment. In these achievement is measured against levels of performance defined by statements of attainment, rather than graded in ways which depend on the performance of other candidates.

One example of these trends is provided by the aftermath of the World Conference on Education for All at Jomtien. This was different from the Universal Primary Education (UPE) initiatives that preceded it in the 1960s (e.g. the Addis Ababa conference). For the first time the declaration included a concern for levels of achievement alongside renewed ambitions to universalise enrolments over the basic education cycle. The implication was that most national systems needed to develop their own country-based definitions of acceptable levels of achievement in different curriculum areas. An important inference was that more stress on criterion-referenced approaches to assessment would assist this process.

This chapter is in three parts. The first part discusses recent thinking on criterion-referenced assessment and raises a number of issues concerned with the advantages and disadvantages of the approach. The second part illustrates some of the dilemmas that may arise by presenting brief case studies of recent developments in the UK and in Mauritius both of which have attempted to link public assessment with criteria for attainment. The last part addresses some issues concerned with the effects of examination practices on internal efficiency with a special concern with repetition rates.

Issues in Criterion-Referencing

Context and Definitions

Discussion of norm-referenced and criterion-referenced testing has a long history. The first mention of approaches to assessment which appear to be criterion-referenced can be traced back into the nineteenth century though it was not until the 1960s that the term was first widely used and an extensive professional literature began to develop. The modern origin of criterion-referenced testing probably lies in the experiences of the U.S. military in 1939-45 in training personnel as operatives of new equipment.»Criterion performance requirements«(Miller 1962)¹ came to specify outcomes and define training experiences. Significantly, most of the tasks to which the approach was applied were closed-ended and often single-event orientated (see below).

¹ For further references see 2.4. Bibliography.

The first wave of emphasis on criterion-referencing techniques in the curriculum development literature was concerned with the use and abuse of behavioural objectives in providing a basis for systematic curriculum development and in defining a framework for teaching and the assessment of learning outcomes. Most recently the growth of various types of vocational provision in many countries has led to increased emphasis on competency-based approaches to training which have been seen as especially suited to criterion-referenced assessment. In a way this echoes the earliest developments of criterion-referencing in relation to military training. Increased stress on criterion-referencing may also reflect the pressures on many education systems to be seen to be publicly accountable in performance terms, and to respond explicitly to demands for increased quality and relevance. Criterion-referencing seems to offer a method to satisfy these developments.

The simplest definitions of norm- and criterion-referenced tests distinguish between tests which are designed to compare the performance of individual students with that of other students, and those which assess performance against a criterion independent of the performance of others. This sounds a simple and powerful distinction. A moment's thought indicates that things are not so simple.

In fact, this common kind of definition does not identify two different types of test very effectively; criterion-referenced test items may look very similar to norm-referenced items. If there is a useful distinction it is more concerned with differences in the interpretation of performance on test items than in the nature of the items themselves.

We should note that tests constructed to produce norm-referenced outcomes can be regarded as criterion-referenced - the selection of particular items creates the criteria.

Similarly, criterion-referenced instruments can be used to rank candidates and compare their performance and may therefore be treated as norm-referenced instruments. In either case it can be argued that the validity, reliability, and suitability to purpose of a test may suffer if it is designed for one purpose and used for another. However the point stands that it is the use made of results, rather than the nature of items that most clearly distinguishes between test types.

The reality is that it is often difficult to find examples of either type of examining which do not contain some elements of the other. For example, public school examinations are often considered to be norm-referenced tests with a certain proportion of candidates placed in each grade, year after year, identified from scores standardised onto a normal distribution. Analysis of results over time often produces patterns inconsistent with true norm-referencing. In principle the proportions of candidates achieving each grade should remain constant from year to year. In practice it is not uncommon to observe that greater proportions of higher grades are awarded if time series data is analysed for some public examinations. Where this is the case the strict assumptions of norm-referencing are broken. It is often thought reasonable to suggest that if actual levels of achievement are judged to be increasing greater proportions should get higher grades. But this argument admits the influence of an underlying attachment to a standard of achievement independent of the population taking the test, which is a kind of criterion. Pass rates between subjects may differ significantly. In a norm-referenced system it is not obvious why this should be. The most obvious explanation is that there is a sense in which different standards are being applied to different subjects.

As noted above the items that are selected for a norm-referenced test can be regarded as defining criteria against which performance is being assessed. Each will be drawn from a notional set of similar items designed to assess similar outcomes. It is difficult to argue against the view that many tests that claim to be designed within norm-referenced assumptions adhere implicitly to judgement of standards which link performance to that in previous years (a historical criterion-reference point) and to the judgements of experts concerning acceptable expectations of levels of performance defined by competence in the subject area. They thus manifest some characteristics of criterion-referencing.

Performance on criterion-referenced tests is supposed to be defined independently of the performance of others. It is also often linked to the concept of mastery, which is a quality

defined dichotomously - either it is possessed or it is not. Successful performance against a criterion is judged to indicate acquisition of knowledge and skills and readiness for subsequent learning. But few real school tests appear to be tests of mastery which simply assess performance against a criterion in terms of success and failure. This is because it is often not possible to define a criterion with sufficient precision to decide unambiguously whether or not it has been achieved. Judgement intervenes and is often influenced by what are thought to be reasonable expectations of performance at a given level with a given group of candidates - a norm-referenced consideration. Moreover, it is unusual to have criterion-referenced assessments that have no gradation in performance beyond an indication of whether or not a particular criterion has been achieved. More often information is provided and sought on the extent to which performance against a criterion may have failed or succeeded. Grades of performance begin to appear on scales. Though there may be fixed points defined by criterion, interpolation may be norm-referenced. Even the fixed points may be defined more by the proportions of candidates they identify than by absolute levels of achievement which is again a process with a norm-referenced flavour.

The distinction between norm- and criterion-referenced testing is therefore blurred and dogmatic approaches to differentiating between the two have not proved particularly helpful. A less rigid approach, which recognises that the approaches are not in opposition, but inextricably linked in practice, can offer useful insights and suggest new procedures.

The rest of this chapter explores some of the key issues related to developing and using assessments that have a criterion-referenced flavour though they may have hints of a norm-referenced pedigree.

The Origins of Criteria

Much is written about the specification of criteria and surprisingly little about from where criteria originate. This is surprising since it is hardly a trivial issue. A first analysis suggests that there are a number of possibilities which include:

- expectations of society concerning competent behaviour,
- the logic of a subject area,
- the judgements of practitioners within a field,
- historically determined expectations of achievement,
- politically determined priorities applied to education.

Criteria may be suggested by collective expectations of competent behaviour. Thus there is likely to be an expectation that an electrician can wire a house or a plumber can mend a leaking tap. These kinds of expectations may not be sufficiently precise to define a criterion operationally, but they may have a strong formative influence.

What is commonly believed about the characteristics of competence is easy to adopt when developing criteria which may be applied to certification since it may require little further legitimisation. It will be widely understood and accepted.

Another possible source of inspiration is the logic thought to be associated with a subject area of the curriculum. A commitment to an empirical approach to the teaching of Science suggests that the knowledge and skill necessary to design experiments, collect data and interpret results should be the basis for the definition of some appropriate performance criteria. Necessary sequencing of ideas may also indicate both the form and ordering of learning criteria (concepts of mass and volume are necessary to comprehend density).

Those who practise a profession may be identified as the group most likely to be able to define the nature of competent behaviour. Criteria for educating scientists could be derived from the opinions of the community of professional scientists. It may need careful consideration to decide whether such experts really are in a position to determine criteria appropriate for the science education of the majority, who will not become professional scientists.

Criteria identified often stand in a clear relationship to previous practice. It is rare that attempts to develop criterion-referenced assessment completely replace existing expectations of levels and types of achievement. More often they refine and extend them.

Thus attempts to introduce criterion-referenced standards to GCSE (General Certificate of Secondary Education) have been compromised by explicitly linking a reference point to previous GCE standards.

Finally, we should not forget that criterion definition may attract the interest of political process. This may seek to include or exclude particular outcomes and incorporate more or less workable criterion statements. It may happen in Science, though not perhaps as much as in other subjects. Nevertheless, ample examples can be found where this has been a consideration.

Curriculum developers and examiners are missing from this list. This is not because they are marginal to the process. It is to act as a reminder that their most valuable role may be to act as an honest broker between criteria emerging from different sources rather than as another vested interest. Durable criteria will not come only from those charged with the technical task of generating working definitions.

Types of Criterion-Referencing Statements

Criterion-referenced statements can usually be located within domains which define the boundaries within which judgements are to be made about achievement. Black and Dockrell classify domains into a threefold typology - single act, closed and open.

A single act domain»comprises a discrete single phenomenon such as the ability to jump 1.25 metres using the straddle jump, or to recall the valency of oxygen«(Black & Dockrell 1984, p. 52). Outcomes are unambiguous. Their statement is often coincident with the appropriate assessable outcome. They may often be judged educationally trivial since they are so narrowly defined and cannot easily be inter-related. They generally stand as discrete achievements.

Closed domains are those cases where it is possible to specify all possible items that could be used to test the domain. For example, if recall of the names of the planets were to be tested all possible cases would be known, though all might not be used in a test that sampled from the full set. The domain is closed because all instances that define acceptable performance are known and can be defined.

Open domains are the most difficult to specify. They are also the ones most commonly applied to educational rather than training tasks. In these cases it is not possible to identify all possible cases that would indicate that performance within the domain was acquired. For example, if it was desired to test whether a concept like gravitational attraction had been understood, there is no way of defining all the potentially infinite set of cases that could be presented that would illustrate mastery of the idea. Open domain specification is more receptive to outcomes that have an integrative character and/or may be defined at a higher level of generality. The price is increasing ambiguity.

The definition of these three types of domain is complemented by another distinction which is of great significance (Black and Dockrell 1984). **Explicit** domain definition occurs when it is possible to specify a domain in advance of constructing items to test its acquisition. This is most likely to be possible in the case of single act domains where the domain virtually specifies the test item. Most discussion of criterion-referenced test construction assumes explicit domain definition as the first step. Satterly (1989) is one of many who adopt this approach. Though this is conceptually attractive it often proves difficult to define domains operationally without developing both domain statement and exemplars.

Implicit domain definition abandons the notion of pre-determined domain statements in terms of a more emergent process. In this case experienced teachers and examiners develop items which in their judgement distinguish between the performance of candidates. The domain is

then defined with reference to these items which are judged to discriminate between candidates that are competent and those who are not. This kind of **instrument defined** domain usually stops short of formal statements that can be scrutinised and compared. It seems to reflect what practitioners often do in developing criterion-referenced instruments. In reality it is the exemplars rather than the domain specification that become definitive.

The Precision of Specification

It is important to consider how precisely criteria can be defined. Throughout the 1960s and 1970s arguments raged about the value of different types of statements of educational objectives. The programmed learning movement in particular attempted to pre-specify every learning outcome on the way to more general educational goals. These attempts to define the outcomes of learning more and more precisely became entwined with the development of criterion-referencing. Familiar Tyler-style objectives (specified in terms of behaviour and content) were elaborated into Mager-style objectives (behaviour, conditions, and standards) and Gagne's five stage objectives (action, object, situation, tools and other constraints, capability to be learned). Most recently in the United Kingdom National Vocational Qualifications (NVQ) specifications typically involve statements of performance criteria, range statements, knowledge and skill requirements, and indications of specific assessment tasks.

Some resisted these developments despite their promise of a precision of definition that would lead inexorably to desired outcomes. They argued that much that was valued could not be captured by behavioural educational objectives and that other types of outcomes (e.g. those located by expressive objectives as discussed by Eisner (1968)) were at least as important. Others began to realise that the search for greater and greater precision in specification was producing diminishing returns and undermining viability. It became clear that, except for the simplest single outcome learning tasks, comprehensive specification of learning outcomes quickly led to the generation of very long and unwieldy lists of behavioural objectives.

Popham wrote one of the first collections of papers on criterion-referencing in the early 1970s (Popham/Baker 1970) and initially identified himself as a committed proponent of the approach. Like others he became concerned that the original idea - that outcomes could be precisely specified - was not achievable. By 1984, he recanted on some of the more extravagant claims for criterion-referencing. Thus he discarded the idea that items that were equivalent but different could be created (functional homogeneity) if only the specification was precise enough, by saying that »About the only way we can ever expect to attain functional homogeneity is to keep pruning the nature of the measured behaviour so that we're assessing ever more trifling sorts of behaviour. That would be inane« (Popham 1984, p. 39, cited in Wolf 1993).

A softer view gained ground that meso-level specification was essential and that interpretation of these through »common understandings« (which involved judgement based on experience and example) was in practice the only workable approach. Wolf (1993) illustrates this by showing how difficult it is to understand the nature of performance specified by detailed statements of behavioural objectives of manageable length, and how much easier this becomes if examples of typical assessment tasks are given. It is usually the examples, rather than the statement of the objective, that are definitive in shaping the items that are produced by those who do not write the criterion statements.

Levels of Acceptable Performance

Acceptable levels of performance can be defined in a number of ways. It may be sufficient to simply assess performance against a stated criterion which is sufficiently precise to define an acceptable outcome, e.g. can 100 metres be run in less than 14 seconds on a single occasion? More often outcomes cannot be so precisely defined. It may also be necessary to sustain performance across several cases to have confidence that special conditions did not apply or that serendipity has intervened in a particular instance.

One interpretation of a criterion-referenced approach is that performance can be judged in

terms of a dichotomous judgement - either the criterion is achieved or it is not. This implies it is really only judgements at the borderline that are important and that it does not matter by how much a performance criterion is exceeded or failed. Whilst this may seem reasonable and workable for individual items, it is usually the case that performance on several criteria is grouped together and these may also measure performance across more than one domain. It begins to appear unreasonable to insist that all items have to be answered successfully in order to demonstrate acceptable performance, though this is sometimes advanced as a condition. Lower thresholds - e.g. 80% of items answered correctly begin to be applied in practice. Whenever this is permitted decisions have to be made on acceptable methods of aggregation.

In many situations value is attached to gradations of performance, partly because it is recognised that the quality of a performance against a criterion usually cannot be reduced solely to statements of measurable outcomes. Assessors and teachers often distinguish between satisfactory and good performance (e.g. on project work) - though two projects may both be considered a pass one may be thought to have reached a different standard than another. Some qualities (for example »elegance«) may be valued in the design of an experiment or the solution of a mathematical problem. These qualities may be felt to warrant recognition and require some grading of the quality of a pass. A level of acceptable performance may then acquire a »flavour« (»starred«, »distinction«) over and above its satisfaction of a criterion.

Judgements about acceptable levels of performance for mastery against criteria may be influenced by norm-referenced considerations. A decision has to be taken as to what it is reasonable to expect as an outcome for a particular group of candidates. It would seem pointless to set criteria of performance at levels unlikely to be achieved by most candidates of a particular age and ability. Similarly, setting criteria at a level regarded as trivial and easily met by most candidates would also be of little value.

Modes of Assessment

The range of options in choosing modes of assessment is very wide. They include individual and group testing; written, oral and practical tasks; open and closed book conditions; self, school-based, or external assessment; continuous formative or infrequent summative assessment; time constrained or unconstrained testing. It is well established that different modes and different formats affect candidates' performance. In the United Kingdom girls appear to perform less well on objective test items and appear to take more time to complete them. Some research suggests that there are cultural differences in trade-offs between speed and accuracy which result in different performance levels in timed tests.

An emphasis on criterion-referencing can be applied to virtually any mode. The format chosen may constrain the range of criterion against which performance can be evaluated. This may be of particular concern where curricula goals are essentially open-ended and assessment instruments are pre-disposed towards closed ended outcomes. A case in point is the widespread reliance on multiple-choice items for testing Primary Science achievement. Though these may be capable of assessing a much wider range of outcomes than they are commonly directed towards, valued outcomes (e.g. manipulative skills, powers of expression, affective outcomes) may be inaccessible through this particular method.

Continuous assessment is as susceptible to criterion-referenced approaches as other forms of examining. It offers the opportunity to reach corners of cognitive and affective achievement that other methods cannot easily reach if a wide range of approaches to gathering assessment data is used. Some recent research has explored possible relationships between external examination results and internal test performance in a sample of schools in Papua New Guinea. Though not conclusive, this study is suggestive that the schools with the highest scoring students have the lowest correlations between external examination score and internal test scores. Conversely the lowest scoring schools appear to have the highest correlations. One possible explanation is that higher performing schools do indeed assess a different and wider range of traits in internal tests than do low performing schools. Clearly the mode of examining can create a different pattern of results (Ross 1994).

Assumed Antecedents

The definition of appropriate criterion-referenced statements generally contains assumptions about antecedent conditions. The specification of a particular learning outcome usually presupposes that other conditions have been met. At the lowest level it will usually be taken for granted that pupils possess adequate levels of literacy and numeracy for them to achieve the criterion specified. By extrapolation it will be assumed that necessary prior learning has taken place and that essential knowledge and skills have been retained.

There is a sense in which the logic of criterion-referencing suggests that all assumed antecedent conditions should be made explicit. They can then be transformed into criteria which would define readiness to approach a new criterion performance level. Though logical, this is of course impractical. The best that may be possible will be to note any essential pre-conditions that might exist that could prevent the possibility of a criterion level of attainment being achieved.

The creation of criterion statements of performance is a forward looking act. The probability of their achievement cannot be separated from previous attainment. This is why the often neglected, but centrally important question of antecedent conditions should be an integral part of attempts to specify criteria.

Reporting Outcomes and the Aggregation Problem

Reporting the results of criterion-referenced tests presents a number of dilemmas. Some of the most important are noted below. First, simply reporting performance in terms of mastery or non-mastery has several disadvantages. It does not distinguish exceptional candidates from those whose passes are marginal; it may allow little insight into the causes of failure or success; single pass/fail boundaries may be so high as to discourage most candidates or so low as to fall into disrepute; pass/fail boundaries may come to be regarded as maximum rather than minimum levels of achievement; pass levels may be achieved through different aggregations which undermine any single construct of mastery.

The last issue merits expansion. Problems of aggregation permeate criterion-referenced and competency-based assessment systems. What at first sight may appear as a simple problem may be very complex to unravel. Once a judgement on performance is compiled from several sources of data the rules associating these have to be determined. What these rules are will determine which candidates meet the criterion and which do not. The kind of dilemmas that arise concern questions of how raw scores on the same criteria should be aggregated; what conditions apply across assessments of different criteria when these are associated together into an overall judgement (are they standardised, are they weighted, do all have to be passed above the criterion level?).

An illustration of the kind of problems that persist concerns the extent to which compensatory performance is permitted. This refers to the degree to which higher performance against one criterion might compensate for low performance against another.

The simple criterion-referenced view of competency - that all its elements can be specified and all essential parts have to be mastered - proves difficult to work in practice. If it is applied it may result in lowering criterion levels sufficiently to allow acceptable proportions of students to be judged competent. In practice some form of compensation is usually allowed. Aggregation rules allow low performance in some areas if the overall performance is judged adequate. But, paradoxically, this accommodation which is necessary to have a workable criterion-referenced system, introduces ambiguity into the meaning of competence or mastery. Those succeeding will have different performance profiles. Mastery will have plurality of meanings.

Criterion-Referencing in Practice - Two Case Studies

England

There are many sources which discuss assessment in the United Kingdom and which relate to the national curriculum and the development of criterion-referencing. This brief review draws attention to some features that may have wider relevance.

The underlying model of attainment is based on the identification of 10 levels of achievement which are specified in relation to attainment targets (ATs) specified for each subject. There are seventeen ATs for Science which are grouped as shown below.

- **Exploration of science, communication and the application of knowledge and understanding.**

AT1 Exploration of Science

- **Knowledge and understanding of science, communication, and the applications of science.**

AT2 The Variety of Life
AT3 Processes of Life
AT4 Genetics and Evolution
AT5 Human Influences on the Earth
AT6 Types and Uses of Materials
AT7 Making New Materials
AT8 Explaining How Materials Behave
AT9 Earth and Atmosphere
AT10 Forces
AT11 Electricity
AT12 The Scientific Aspects of Information Technology
AT13 Energy
AT14 Sound and Music
AT15 Using Light and Electromagnetic Radiation
AT16 The Earth in Space
AT17 The Nature of Science

Examples of Attainment Targets are included in the appendix. **AT1 Exploration of Science** is concerned with the development of intellectual and practical skills that allow the exploration of the world of science, the development of understanding of scientific phenomena, and familiarity with the procedures of scientific exploration and investigation. Broadly speaking they cover the general skills associated with:

- Planning, hypothesising and predicting.
- Designing and carrying out investigations.
- Interpreting results and findings.
- Drawing inferences.
- Communicating exploratory tasks and experiments.

The remaining ATs are all classified under knowledge, understanding, applications and implications of science. AT2 is concerned with the variety of life and focuses on knowledge and understanding of the diversity and classification of life and of the relationships, energy flows, cycles of matter and human influences within ecosystems.

There are four **Key Stages (KS)** in the national curriculum. KS1 covers ages 5-7 approximately, KS2 8-11, KS3 12-14, KS4 15-16.

For Science KS1 and 2 include ATs 1-6 and 9-16 only. The ten levels of performance are related to Key Stages as follows:

- KS1 1-3
- KS2 2-5
- KS3 3-7
- KS4 4-10

Each level of performance for each attainment target has a statement defining the level. The diagram shows how the Key Stages relate to the range of attainment targets at different ages.

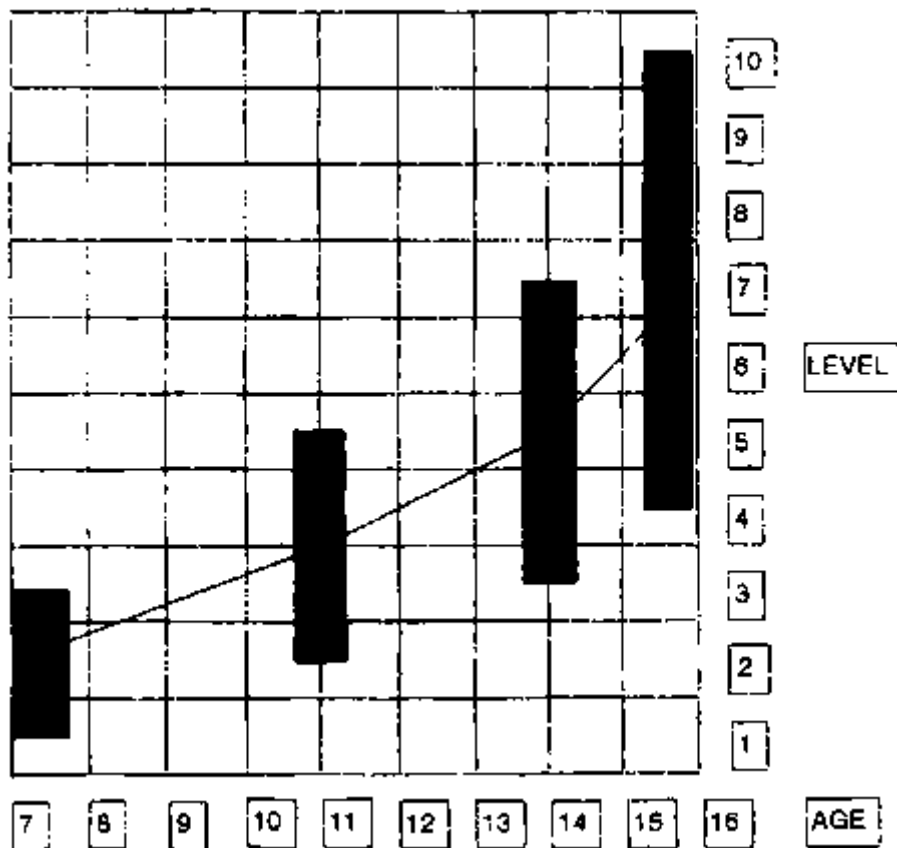


Diagram 1: PUPIL ACHIEVEMENT LEVELS - TGAT MODEL

The expectation is that average pupils will reach level 2 by the end of Key Stage 1 (KS1), level 3/4 at the end of KS2, level 5 at the end of KS3, and levels 6/7 at the end of KS4. Achievement will be spread over a range of levels since pupils will differ in their ability and attainment at any particular age. Progression through the levels is anticipated for all children. Pupils will have the psychological reward of reaching higher levels as they get older, unlike in norm-referenced systems.

Important issues which have arisen in implementing this system are summarised below.

- (1) The national curriculum has introduced statements of attainment that define levels of performance on the Science curriculum for each domain of attainment. This is the first time there have been a single set of outcomes defined for Science teaching throughout the primary and secondary school. The specification is essentially criterion-referenced in the sense that the attainment targets (ATs) represent the criteria. However, it is not clear that the ATs are defined sufficiently precisely to ensure what is expected is the same in every school.

(2) From the outset a range of levels of achievement has been anticipated from children of a given age/grade. Performance is reported in terms of level and no one level is the only one appropriate to children in a particular grade. Thus normal children are not all expected to reach say level 5 at one particular age - some will work at level 4, some at level 5 and some at level 6. The range of performance considered normal increases for the higher Key Stages. This recognises that differences in achievement may have a cumulative character that heightens differences over time. KS4 in Science actually has two variants in the specification of ATs, one for more able pupils and one for those who are less capable. It is not clear whether the step size between levels remains constant or increases. The definition of normal expectations seems to suggest that step size increases with level. No clear expectations exist about the distribution of pupils at the end of Key Stages across levels.

(3) There has been considerable debate about the linking of grades from the previous assessment system (i.e. GCSE) to performance levels. There is a problem of reconciling the expectation of an equivalence in standard to the principle of regarding the levels of attainment as being criterion-referenced. To admit linking to a norm-referenced examination appears to contradict the intention of establishing a criterion-referenced system.

(4) Since it is expected that »typical«pupils should perform at Levels 2, 3/4, 5, and 6/7 at ages 7, 11, 14 and 16 there is a sense in which operationally AT criteria are likely to be interpreted in ways which result in this pattern of achievement. To put it another way, if there was a wide divergence from this pattern the framework of Key Stages, ATs and levels of attainment would be undermined. The issue is clearly addressed in the Dearing Report which discusses linking the mid-point of achievement at the end of each Key Stage to »the actual performance of a random sample of pupils on the knowledge, understanding and skills required by a programme of study within a particular Key Stage«(Dearing, R., 1993, p. 43). This is a clear indication of an intention to norm-reference.

(5) As currently specified, it is not clear what procedures are appropriate for the aggregation of performance on attainment targets (ATs) in different domains, or what conditions relate to satisfactory performance at a particular level within a domain. The former is problematic if compensation is permitted and good performance in one domain can make up for poor performance in another. Can exceptional achievement on AT1 compensate for poor performance on AT2 and 3? Within a domain, performance at a particular level may be demonstrated by a pupil in one assessment context but not in another designed to assess the same attainment target at the same level. What conclusion is to be drawn?

(6) Comparability in levels of achievement at the same level in different Key Stages is problematic. Is it really likely to be the case that a level 5 performance by an 11 year old carries the same meaning as a level 5 performance of a 16 year old. It seems unlikely that this is literally the case though in principle it should be so.

(7) The burden of assessment may be excessive. The average pupil reaching level 7 after 11 school years would have been assessed on some 700 statements of attainment. A classroom teacher at Key Stage 1 (7 year olds) with 35 pupils who assessed all pupils against every appropriate statement of attainment would make and record about 8000 judgements each year. This has struck some as an excessive load that has begun to erode teaching time in favour of assessment. Recent changes have responded to this by reducing the burden by as much as 50% at some levels.

(8) The use of attainment targets to define attainment fragments the curriculum. Those outcomes which involve the integration of a variety of skills and understanding are likely to be undervalued and ignored in the assessment process. Skills employed in design and development of experiments may not readily fall into separately assessable components. The whole activity may be more than the sum of its parts. This problem is a strong argument against overspecification of ATs to more and more precise levels.

Mauritius

The Mauritius Examination Syndicate (MES) is experimenting with the introduction of Essential Learning Competencies (ELCs) and Desirable Learning Competencies (DLCs), building on work originally undertaken at the National Council for Educational Research and Training (NCERT) in India. The philosophy behind this is essentially one of criterion-referencing performance in grades 4 and 5 on school-based tests, and in grade 6 which leads to the Certificate of Primary Education.

Essential Learning Competencies are intended to represent the levels of learning achievement in all subjects (knowledge, understanding, skills, abilities, interests, attitudes and values) which are considered minimum but essential for all students to acquire. It is argued that they can be regarded as attainment targets below which learning achievement is not sustainable. Since there is a range of ability and children will progress at different rates, higher order competencies have been specified involving more complex mental processes and/or learning content. These are denoted Desirable Learning Competencies (DLCs) and are treated as optional levels of achievement for certification. Performance on DLCs are considered, in combination with that on ELCs, for selection into secondary schools. This is thus a kind of two level system, though currently the relationships between the levels are not specified in any particular way.

The Mauritius Examination Syndicate has recognised the impossibility of measuring competencies directly in favour of defining detailed learning objectives. Performance on these can be assessed through observable behavioural outcomes. Each subject in the Certificate of Primary Education Examination (CPE) was analysed using a simple taxonomic model (Knowledge, Understanding, Application for »content subjects« and Comprehension and Expression for languages) to identify groups of objectives that could be specified behaviourally. These groups were discussed with a large number of teachers to validate their classification. The lists of »Learning Competencies« that emerged were judged against six criteria. These were:

1. **Sustainability** - achievement that would be retained and used at subsequent levels.
2. **Communicability** - to ensure that each statement had a clear meaning using a standard presentation including two dimensions (content, ability).
3. **Learning Continuity** - to relate ELCs/DLCs to those that come before and after as part of a continuum.
4. **Functionality** - to ensure that statements are at an appropriate level of generality.
5. **Measurability** - to ensure that a learning competency was measurable with a well defined range of acceptable outcomes.
6. **Achievability** - to adjust ELCs/DLCs to appropriate contextual factors including the developmental level of pupils.

As the system is implemented ELCs/DLCs will be used to structure the CPE and should provide diagnostic insights into performance in grades 4 and 6.

Some examples of ELCs and DLCs are included in the appendix (see: chapter 4.5.). Important issues with the implementation of this system appear to be:

- (1) ELCs/DLCs are not graded by level. As statements of learning competencies they imply that either candidates possess these competencies or they do not. This may be acceptable on a competency by competency basis but there are obvious risks when they are applied to the population of pupils in a particular grade. ELCs may be set at a low level to ensure that most pupils do in fact demonstrate these competencies. In this case the tasks they present to average and above average pupils are likely to be seen as trivial. Alternatively the ELCs may be set at a higher level to challenge average and above

average pupils. If so large proportions will fail to meet the criteria and they will cease to have practical significance as minimum levels of achievement. The problem lies in specifying a single level rather than a band of acceptable achievement levels for a grade.

(2) There is a risk that minimum levels, wherever they are set, become maxima if there is no recognition of performance above the minimum. If performance is assessed dichotomously then there is no particular incentive for pupils to perform above the minimum level necessary to achieve the ELC or DLC. In so far as this might encourage teachers to give more attention to those pupils having difficulty in acquiring the competency this might be welcomed. If it suppressed the performance of the best pupils it might be a matter of concern.

(3) The problems of aggregation and compensation remain under discussion. Is performance on ELCs simply the total number of ELC items correctly completed? Should minimum scores apply to domains or topics before a competency is deemed to be acquired? Can performance in one area compensate for performance in another either within subjects or across subjects at CPE? How can performance on ELCs be aggregated with performance on DLCs?

(4) Can adequate progression be demonstrated for ELCs and DLCs at different levels? Or are they simply more educational objectives which extend the range but not the level of achievement assessed?

Assessment and Repetition

Table 1 displays a number of indicators of levels of educational development in African countries based on the most recent UNESCO data. From this we can see that unweighted Gross Enrolment Rates (GERs) averaged about 76% in 1990 as they did in 1980. Gender disparities appear prominently for both GERs and Net Enrolment Rates (NERs). At secondary level both GERs appear to have increased over the last decade as, in most countries, secondary school systems have expanded at a faster rate than primary. Average secondary GERs have increased from about 17% to 23%. In most countries transition rates from primary to secondary have almost certainly increased over this period. The average ratio of GER secondary to GER primary increased from 0.23 to 0.28 suggesting that this was a common pattern. Repetition often appears as one of the main sources of internal inefficiency in school systems which results in a suppression of GERs below levels they would otherwise reach (Colclough with Lewin 1993).

Enrolment rates, the rate of flow of pupils through the school system, and the transition rate into secondary are all influenced by assessment practices and levels of achievement. Table 1 shows that repetition can be high in all grades through the primary school system. Characteristically it is greatest in the grade preceding primary school leaving examination levels, as failing students repeat to improve their chances of gaining access to secondary school. This is most commonly at grade 6 and repetition averages 28% at this level. Average repetition rates in each grade are around 20% implying that enrolment rates could be about 20% greater at no extra cost if repetition was reduced to insignificant levels. Repetition is a very significant source of internal inefficiency in many systems.

Internal school assessments generally determine rates of repetition and promotion through the primary school. The information which is used to determine whether or not students repeat will vary considerably from system to system. Most commonly teachers make judgemental decisions on unstandardised data drawing on experience with other students. Although this may appear in some sense to be criterion-referenced (teachers applying the criterion of successful completion of a year's work), in practice it may well have a strong element of norm-referencing, with those pupils well below average in performance being retained in the grade for that reason.

It is also likely that individual teachers will not wish to have repetition rates amongst their classes much greater than other teachers, so a »normal« expectation may exist. In addition there may be national norms (as in China under the nine year basic education law)

good and poor schools by comparison of examination results are deeply embedded in the culture of teachers. Attempts at criterion-referencing do encourage clear consideration of educational purpose which ought to be part of every teacher's lesson planning. There is no particular reason why criterion-referenced instruments should be more time-consuming to construct and use than well constructed norm-referenced assessments. Often individual items within a criterion-referenced set may be similar to well constructed items used for norm-referenced assessment as noted earlier. The real problem is to find ways of improving the reliability and validity of internal school assessments and adopting simple criterion-referencing can help with this.

In countries where there are not enough places in secondary schools to accommodate all those completing primary school the transition from primary to secondary is generally controlled through a primary school leaving examination. These examinations are often retained when transition rates reach high levels and almost all children proceed to secondary since selection (for different secondary schools) remains an issue. An external examination also provides information which is otherwise unavailable which can be used to monitor school performance.

Primary school leaving examinations have a considerable importance in determining life chances. The content and format of them often washes back onto curriculum decisions and assessment practices at the school level. They exercise extensive influence over teaching and learning (Dore 1976, Oxenham 1984). They also determine the amount of repetition that takes place in the transition grades.

Two types of repetition occur in the final grades of the primary cycle. The first is amongst those who have failed a primary school leaving examination and repeat in order to pass. The second type of repetition is by those who have passed but who wish to improve their ranking to be selected by better secondary schools. In some systems (e.g. Mauritius) the latter may reach 40% of all repeaters in the last year of primary.

Where external assessment instruments are norm-referenced and pass and fail grades are determined on a norm-referenced basis, it will be the case that a pre-determined proportion of candidates will »fail«. This will remain true whether or not their actual level of achievement improves. Assessment in this case is a zero sum game where improvement by one child (or school) must be balanced by relative deterioration by another. There will be as many pupils and schools above average as below and not all pupils/schools/districts can be above average (as in the debate concerning whether 50 States in the USA could all be above average on elementary school tests; Cannell cited in Shephard 1991).

In norm-referenced systems it is not clear what implications should be drawn where failure rates (and consequent repetition) is great. The problem is that if the assessment really is norm-referenced then high rates of failure are the result of conscious design and cannot be changed by improved performance amongst candidates. Thus in some systems pass rates in key subjects like English and Mathematics may be as low as 20 -30%. If the assessment is norm-referenced this implies a particular judgement of where to draw the pass/fail boundary. It carries the message that the achievement of the majority of candidates in these subjects is not worth recognition. Pedagogically this seems undesirable. If pass rates vary between subjects in norm-referenced systems it is because of choices made to make some subjects relatively difficult rather than because some subjects are intrinsically more difficult.

Introducing some elements of criterion-referenced assessment to primary school leaving examinations is claimed to be attractive for a number of reasons. In principle it would seem to:

- allow performance to be measured against a fixed and public standard,
- reward improvements in performance since all pupils, schools and districts could see performance improve at the same time,
- permit certification based on criteria,

- encourage educational objectives to be more clearly specified.

The main disadvantage is that criterion-referencing would not necessarily »pass« the numbers of pupils for which places exist at secondary level. It might also make it more difficult to finely grade performance. Leaving qualifications certifying specific levels of achievement related to criteria would not guarantee entrance to secondary schools. They might therefore be seen to have limited value. Significant numbers of candidates might be »qualified« according to entrance criteria but would not necessarily be admitted to a secondary school. A greater emphasis on criterion-referenced approaches could increase the probability of curriculum sensitive item writing, improve the quality of diagnostic information available and improve standards of attainment (Lewin 1992) if introduced flexibly without over specification of attainment levels.

Concluding Remarks

This chapter has discussed the issues that surround the development of criterion-referenced assessment instruments and identified major problem areas. It has summarised recent experience with the national curriculum in England and with Learning Criteria in Mauritius. Finally, it explored some of the relationships that exist between decisions on testing and repetition and transition rates from primary to secondary school.

The title of the chapter contains a question that has not been answered. Is criterion-referencing a panacea that will overcome objections to norm-referenced styles of assessment, or is the new emphasis a false dawn? This analysis suggests that it is neither. Criterion-referencing cannot of itself resolve many of the enduring dilemmas in assessment. Intrinsically it is no more reliable or valid than other approaches. Neither is it more cost-effective or technically less demanding. Nor can it resolve the difficulties of undertaking and legitimating the selection of some students from many. The approach does have the special value that it forces systematic consideration or purpose in a more focused way than most others. For this reason alone its development is worth encouraging. It is more than a palliative that allows prevarication. It can actually help draw attention to critical issues that require policy decisions and may make it more difficult to avoid these. The problem of repetition and its reasons is one area in which this is evident.

In conclusion the challenge for the future is to explore the complementarities between criterion-referenced and norm-referenced approaches to assessment. Neither on their own are sufficient to service the needs for information to aid decision making on individuals, schools, and education systems.

2.2. Test Development - Designing Tests and Presenting Results

Keith Lewin

Introduction

This chapter provides a simple introduction to the construction of some forms of assessment. To design an assessment instrument decisions must be made in three main areas These, are:

- (1) Deciding on the purpose(s) of the test.
- (2) Developing (or utilizing existing) statements of specific measurable learning objectives.
- (3) Selecting the most appropriate test instruments.

The first three parts of this chapter discuss these issues briefly. The fourth part lists some key concepts. The fifth part briefly discusses the construction of simple tests. Part six and seven provide some examples of types of multiple-choice questions and a checklist of good practice in their construction. Finally, the last two sections comment on reporting results and introduce

simple standardization procedures.

Purposes

It is always important to have clearly in mind the purpose(s) for which a test is being constructed. Its nature will depend upon, amongst other things, the uses to which results will be put (rank ordering for selection, diagnostic feedback on individuals or on the effectiveness of teaching, evaluation of curriculum materials, comparison of previous and present achievement of individuals etc.). A recent list of purposes covers the range likely to be encountered (Pennycuik 1990):

Recording and reporting attainment	Diagnosis and remediation
Certification and qualification	Curriculum evaluation
Selection and social regulation	Feedback on teaching and organizational effectiveness
Prediction	Teacher motivation and appraisal
Measurement of individual differences	Curriculum control
Motivating students	Evidence for accountability/resource provision
Monitoring progress and feeding back information	Maintaining or enhancing standards
Guidance	

Assessment plays a central role in teaching strategies; it can provide opportunities for revision (both before and after administration), it may be used to motivate students, its nature almost certainly influences patterns of learning. Assessment is not therefore a necessary chore, but an integral part of teaching and learning. It is thus important that those who use assessments, as well as those who design them, have a clear idea of their purposes and what it is hoped will be achieved through the use of particular assessment instruments.

Objectives

To be able to construct any assessment that measures achievement for whatever purpose it is essential that such achievement is specified in terms of measurable and observable outcomes. A well designed curriculum will indicate the nature of its intended outcomes. If it does not then users must develop such specifications in order to be able to construct an assessment instrument that is valid.

For any curriculum there is likely to be a range of objectives of different types, e.g. at general and more specific levels, short-term and long-term, within different domains (cognitive, affective etc.). Not all of these will be susceptible to direct measurement during the teaching of a course and some perhaps not at all. Though it may be that some desirable objectives are of a kind which cannot be assessed easily many valued outcomes are likely to be directly assessable, albeit with varying degrees of precision. Without systematic attempts to define intentions and measure progress towards their achievement it will be difficult if not impossible to make judgements concerning the effectiveness of teaching.

Several taxonomies of education objectives exist in which attempts have been made to break down learning behaviour into well defined categories. An example of such a categorization in the cognitive domain with descriptions of meaning and illustrative objectives derived from them is included below (see Table 2, derived from Bloom et al. 1971). References to other taxonomies may be found in books cited in the reading list.

Table 2: Major Categories of Cognitive Objectives with Examples		
Descriptions of the Major Categories in the Cognitive Domain	Illustrative General Instructional Objectives	Illustrative Behavioural Terms for Stating Specific Learning Outcomes
<p>1. Knowledge Knowledge is defined as the remembering of previously learned material. This may involve the recall of a wide range of material, from specific facts to complete theories, but all that is required is the bringing to mind of the appropriate information. Knowledge represents the lowest level of learning outcomes in the cognitive domain.</p>	<p>Knows common terms. Knows specific facts. Knows methods and procedures. Knows basic concepts. Knows principles.</p>	<p>Defines, describes, identifies, labels, lists, matches, names, outlines, reproduces, selects, states.</p>
<p>2. Comprehension Comprehension is defined as the ability to grasp the meaning of material. This may be shown by translating material from one form to another (words to numbers), by interpreting material (explaining or summarizing), and by estimating future trends (predicting consequences or effects). These learning outcomes go one step beyond the simple remembering of material, and represent the lowest level of understanding.</p>	<p>Understands facts and principles. Interprets verbal material. Interprets charts and graphs. Translates verbal material to mathematical formulas. Estimates future consequences implied in data. Justifies methods and procedures.</p>	<p>Converts, defends, distinguishes, estimates, explains, extends, generalizes, gives examples, infers, paraphrases, predicts, rewrites, summarizes.</p>
<p>3. Application Application refers to the ability to use learned material in new and concrete situations. This may include the application of such things as rules, methods, concepts, principles, laws and theories. Learning outcomes in this area require a higher level of understanding than those under comprehension.</p>	<p>Applies concepts and principles to new situations. Applies laws and theories to practical situations. Solves mathematical problems. Constructs charts and graphs. Demonstrates correct usage of a method or procedure.</p>	<p>Changes, computes, demonstrates, discovers, manipulates, modifies, operates, predicts, prepares, produces, relates, shows, solves, uses.</p>
<p>4. Analysis Analysis refers to the ability to break down material into its component parts so that its organizational structure may be understood. This may include the identification of the parts, analysis of the relationships between parts, and recognition of the organizational principles involved. Learning outcomes here represent a higher intellectual level than comprehension and application because they require an understanding of both the content and the structural form of the material.</p>	<p>Recognizes unstated assumptions. Recognizes logical fallacies in reasoning. Distinguishes between facts and inferences. Evaluates the relevancy of data. Analyses the organizational structure of a work (art, music, writing).</p>	<p>Breaks down diagrams, differentiates, discriminates, distinguishes, identifies, illustrates, infers, outlines, points out, relates, selects, separates, subdivides.</p>

<p>5. Synthesis Synthesis refers to the ability to put parts together to form a new whole. This may involve the production of a unique communication (theme or speech), a plan of operations (research proposal), or a set of abstract relations (scheme for classifying information). Learning outcomes in this area stress creative behaviours, with major emphasis on the formulation of <i>new</i> patterns or structures.</p>	<p>Writes a well-organized theme. Gives a well-organized speech. Writes a creative short story (or poem, or music). Proposes a plan for an experiment. Integrates learning from different areas into a plan for solving a problem. Formulates a new scheme for classifying objects (or events, or ideas).</p>	<p>Categories, combines, compiles, composes, creates, devises, designs, explains, generates, modifies, organizes, plans, rearranges, revises, rewrites, summarizes, tells, writes.</p>
<p>6. Evaluation Evaluation is concerned with the ability to judge the value of material (statement, novel, poem, research report) for a given purpose. The judgements are to be based on definite criteria. These may be internal criteria (organization) or external criteria (relevance to the purpose) and the student may determine the criteria or be given them. Learning outcomes in this area are highest in the cognitive hierarchy, because they contain elements of all of the other categories, plus conscious value judgements based on clearly defined criteria.</p>	<p>Judges consistency of written material. Judges the adequacy with which conclusions are supported by data. Judges the value of a work (art, music, writing) by use of internal criteria. Judges the value of a work (art, music, writing) by use of external standards of excellence.</p>	<p>Appraises, compares, concludes, contrasts, criticizes, describes, discriminates, explains, justifies, interprets, relates, summarizes, supports.</p>

When devising objectives that utilise verbs related to statements of behaviour it is important to remember that it is the candidates that have to discriminate between meanings. The difference between comparing and contrasting may be clear to test constructors; experience would suggest that often it is unclear to many school pupils. Whatever system of classifying objectives is used some useful guidelines apply if tests are to be developed based on statements of learning objectives.

- Objectives should be stated in terms of student behaviour, not in terms of learning activities or the intentions of the teacher.
- Objectives should begin with an action verb that indicates the behaviour a student is expected to demonstrate.
- Objectives should be stated in terms of observable changes in behaviour.
- Objectives should be stated as precisely as possible in terms that have agreed meanings.
- Objectives should relate to only one process at a time.
- Objectives should be stated at an appropriate level of generality.
- Objectives should represent intended outcomes of learning experiences.
- Objectives should be realistic given the context.

In constructing objectives it is as well to remember that some verbs describe behaviour far

less ambiguously than others (see the list below from Mager 1962).

Verbs open to many interpretations	Verbs open to fewer interpretations
to know	to write
to understand	to recite
to really understand	to identify
to appreciate	to differentiate
to appreciate fully	to solve
to grasp the significance of	to construct
to enjoy	to list
to believe	to compare
to have faith in	to contrast

Selection of Instruments

Having decided on the purpose of a test and arrived at appropriate statements of observable outcomes, a strategic choice can be made between available testing techniques. Four relevant background questions should help to narrow the choice. These are:

- (1) Is the technique viable? (e.g. Does it demand skills and staff-time in setting and marking that are unavailable?).
- (2) Can it measure effectively the behaviours we are interested in? (e.g. Multiple-choice questions cannot test powers of expression).
- (3) Is it appropriate? (e.g. Does it reinforce or undermine overall objectives? -A testing strategy concentrating on recall of information will not encourage the application of concepts and principles by students).
- (4) What assumptions do we make in opting for a particular strategy? (e.g. Are students sufficiently literate to understand and reply to test questions or is the test primarily testing literacy?).

The range of available techniques is very wide. Some of the common possibilities are indicated below. Most combinations from the columns are possible.

Modes	Methods	Conditions	Marking	Occurrence	Timing
Individual	Written	Open Book	Self	Continuous	Fixed Time Limit
Group	Oral Practical	Functional Information Closed Book	Internal External	Frequent Infrequent	Flexible Time Limit

Questions in written examinations commonly take two main forms, i.e. **Open Response** where a candidate has freedom to structure an answer in whatever way he/she feels is appropriate (typically »essay« questions); **Closed Response** where questions are structured to admit only one unambiguously defined response (typically objective multiple-choice questions). Between these two extremes lie many different types which use different levels of structure and varying degrees of pre-specification of acceptable responses.

	Open Response	Closed Response
Abilities measured	Powers of expression can be tested. High levels of reasoning may be required, e.g. inference, organization of ideas, comparison and contrast, synthesis, evaluation. Inefficient at indicating knowledge of factual information.	Powers of selection of correct response from given options - may work by elimination. Recognition may be more important than recall. Individual words or phrases may be required for recall. Can operate at high reasoning levels, though synthesis and evaluation difficult to test. Measures factual knowledge efficiently.
Scope	Covers a limited area, questions can usually only sample small parts of curriculum. Fluent students can avoid displaying the extent of their ignorance.	Curriculum content can be efficiently covered at knowledge levels. Reliability benefits from this.
Motivational Implications	Encourage development of powers of expression and the ability to organize material in a communicable form. Selective in-depth consideration of curriculum worthwhile.	Encourages the development of broad background knowledge and abilities.
Examiner	Small numbers of questions therefore less preparation.	Large numbers of questions, many pitfalls in their construction. Time-consuming and demanding.
Scoring	Time-consuming and unreliable, allows comment on candidates' reasoning.	Rapidly scored and reliable.

Some Key Concepts

There are a number of terms associated with tests that it is necessary to be familiar with. Amongst the most important are:

- Validity

This has two primary meanings; **Content Validity** is the degree to which a test comprehensively samples achievement of the objectives of a course in relation to its content; **Predictive Validity** is the extent to which test results can be relied upon as an indication of future performance. Content validity can be applied at individual question level.

- Reliability

This has many definitions related to different ways in which it can be calculated. It refers to the confidence that can be placed in achievement on a test (or individual items) as being a replicable indication of performance (e.g. would the results be the same if marked by different examiners or if taken under similar conditions by the same candidates at a different time).

- Facility

An indication of how easy a test (or item) is for average candidates (or conversely called Difficulty, defined by how difficult a test (or item) is for average candidates).

- Discrimination

An indication of how effectively a test (or individual item) discriminates between the performance of high and low achieving candidates.

- Norm-Referenced Tests

Tests where the performance of a candidate is compared to the performance of other candidates, e.g. a public examination where a fixed proportion of candidates achieve particular grades.

- Criterion-Referenced Tests

Tests where the performance of a candidate is expressed in terms of its relationship to a fixed criterion that is dependent on the nature of the achievement being measured, not the performance of others.

- Domain-Referenced Tests

A test consisting of items which sample representatively performance in a clearly defined area of knowledge or skill.

- Table of Specifications

This refers to a test plan, usually in the form of a grid, that tabulates different characteristics of a test, e.g. content areas and behavioural outcomes for each item in a test (see below).

Topics	Abilities/Behaviours:				
	Knowledge	Comprehension	Application	Analysis/ Synthesis Evaluation	Total %
1. Solid Friction	1,2,3	4	5		25
2. Surface Tension	6,7,8,9	10		11	30
3. Elasticity	12,13	14	15		20
4. Momentum	16,17	18	19	20	25
Total %	55	20	15	10	100

Specifications of this kind may be useful in indicating the distribution of items and planning the desired coverage. Information on item characteristics may also be included (e.g. facility value). A modified grid of this kind can be used to keep track of individual pupil progress and achievement throughout a year.

Simple Test Construction

In constructing simple tests a number of stages are desirable. These clearly vary depending on the purpose(s) identified and the strategic decisions made about the instruments to be used. By way of example let us consider the design of a multiple-choice question paper. Though different decisions may be made on the inclusion of items depending on whether the test is to be treated as norm- or criterion-referenced, much of the procedure likely to be used has common elements.

Sequence of Events

(1) Make decisions on purpose and identify statements of testable objectives. If the test is criterion-referenced identify domains which will be assessed.

(2) Map a **Table of Specifications** for the test. Greater specificity of the table may make it easier to decide whether a question is appropriate for inclusion. The map should indicate how domains are covered by items.

(3) Write test questions (see checklist of Use in Constructing Multiple-Choice Questions). Domain definition may suggest criterion-referenced items.

(4) Logically review test question items - preferably getting another experienced member of staff to participate. This may include:

- consideration of the question as a whole; e.g. for validity, relevance, likely facility;
- checking classification against a Table of Specifications;

- answering questions without referring to answers anticipated;
- examining questions in terms of »checklist«criteria;
- ensuring that presentation and format allow for ease of understanding, response and marking.

(5) Empirically review test items:

- item facility
- item discrimination
- reliability assessments
- instructional sensitivity procedures.

Illustrative examples of types of items, a list of some common pitfalls in their construction, and a Checklist of points related to writing items are provided on the next pages.

Some Examples of Item Types

1. Free Response

How far is it from Cairo to Nairobi? When will the next election take place?

2. Fixed Response (Multiple-Choice)

• What is the capital of Zimbabwe? (knowledge of specific facts)	A	Colombo	(STEM)
	B	Lagos	
	C	Lilongwe	
	D	Accra	
	E	Harare	

What is the product of 5 and 3? (problem solving)	A	1 2/3
	B	8
	C	15
	D	35
	E	112

• If an electric refrigerator is operated with its door open in a perfectly insulated sealed room, what will happen to the temperature of the room? (prediction)	A	It will rise slowly.
	B	It will remain constant.
	C	It will drop slowly.
	D	It will drop rapidly.

• Should merchants and middlemen be considered as producers or non-producers. Why? (decision, explanation)	A.	As non-producers, because they make their living of producers and consumers.
	B	As producers, because they are regulators and determiners of price.
	C	As producers, because they aid in the distribution of goods and bring producer and consumer together.
	D	As producers, because they assist in the circulation of money.

• A rubber balloon inflated with hydrogen is released from the earth's surface. As it rises it increases in size and finally bursts. This is because:

- (application)
- A The temperature inside the balloon increases.
 - B The pressure of the air outside the balloon decreases.
 - C The air outside gradually enters the balloon.
 - D The increase in the temperature of the air outside the balloon causes the balloon to expand.
 - E The changes in temperature outside the balloon weaken the rubber.

3. Clarification/Matching Pairs

Choose the letter heading which most appropriately answers the question. Each heading may be used once, more than once, or not at all.

- | | | |
|---|------------|---|
| A | Aluminium | 1. Which one is a non metal... |
| B | Copper | 2. Which one is normally stored underwater... |
| C | Phosphorus | 3. Which one is the most electropositive... |
| D | Sodium | 4. Which one forms a black oxide... |
| E | Zinc | |

4. Multiple Completion/Selection

Which of the following are insects?

- | | | | | |
|---|---------|------------|----------|---------|
| | 1. ants | 2. spiders | 3. flies | 4. mice |
| A | 1 2 3 4 | | | |
| B | 1 2 3 | | | |
| C | 1 2 4 | | | |
| D | 1 3 4 | | | |
| E | 3 | | | |

5. Proper Sequence

Which one of the following lists the animals in order of the size of normal adults?

- A elephant, cow, rabbit, goat
- B cow, elephant, rabbit, goat
- C elephant, cow, goat, rabbit
- D cow, elephant, goat, rabbit

6. True/False

Kenya borders on:

- | | | | |
|---|----------|--------------------------|--------------------------|
| | | True | False |
| A | Malawi | <input type="checkbox"/> | <input type="checkbox"/> |
| B | Zaire | <input type="checkbox"/> | <input type="checkbox"/> |
| C | Tanzania | <input type="checkbox"/> | <input type="checkbox"/> |
| D | Uganda | <input type="checkbox"/> | <input type="checkbox"/> |

7. Assertion/Reason

A ship floats with a smaller volume submerged in sea water than in fresh water when carrying the same load *because*

the density of sea water is greater than that of fresh water.

- SELECT
- A If both statements are true and the second is a legitimate explanation of the first.
 - B If both statements are true and the second is NOT a legitimate explanation of the first.
 - C If the first statement is true and the second false.
 - D If the first statement is false and the second true.
 - E If both statements are false.

8. Attitude Measuring Questions

Most commonly Likert-type scales (with statements and four of five responses of the agree-disagree type) are used though there are alternative techniques, e.g.

(1) Most of the things I learn in Science lessons I find interesting.

- Strongly agree
- Agree
- Undecided
- Disagree
- Strongly disagree

(2) I like Science practicals more than Science theory periods.

It is good practice to include both positive and negative statements relating to the same idea to provide a balance in the format of items. It is also preferable to have several items on the same theme so that it is possible to form a scale for a particular attitude by aggregating responses. This is likely to provide a more reliable measure than single items.

9. The following items are presented for criticism

1. How many colours are there in the rainbow?

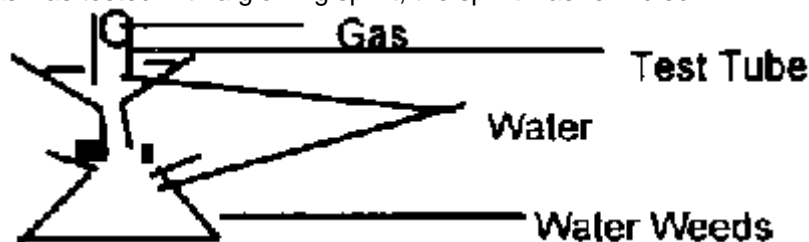
- | | |
|---|----|
| A | 1 |
| B | 3 |
| C | 5 |
| D | 7 |
| E | 15 |

2. Archimedes' principle states that

- A For every action there is an equal opposite reaction.
- B Objects less dense than water float.
- C Plug vortices are anticlockwise in the Northern Hemisphere.
- D Gold is heavier than silver.
- E When a body is immersed in a liquid the upthrust is equal to the weight of liquid displaced.

3. The apparatus shown below was set up and after a few hours in sunlight, it was observed that some gas was collected in the test tube.

When the gas was tested with a glowing splint, the splint was rekindled.



This experiment shows that in sunlight the water weed

- A Produces carbon dioxide
- B Produces oxygen
- C Produces hydrogen
- D Respires
- E Transpires

4. Physiology teaches us that

- A The development of vital organs is dependent on muscular activity.
- B Strength is independent of muscle size.
- C The mind and body are not influenced by each other.
- D Work is not exercise.

5. In the definition of a mineral which of the following is incorrect?
 A It is produced by a geological process.
 B It has distinctive physical properties.
 C It does not contain more than one element.
 D Its chemical composition is variable.

6. Meat can be preserved in brine because
 A Salt is a bacterial poison.
 B Bacteria cannot withstand the osmotic action of brine.
 C Salt alters the chemical composition of the food.
 D Brine protects the meat from contact with the air.

7. What does the term growth mean?
 A Maturation
 B Learning
 C Development
 D Gauche
 E None of these

8. What is the best definition for a vein?
 A A blood vessel carrying blood going to the heart.
 B A blood vessel carrying blue blood.
 C A blood vessel carrying impure blood.
 D A blood vessel carrying blood away from the heart.

Checklist of Use in Constructing Multiple-Choice Objective Questions

(see Ebel/Frisbie (1991) *Essentials of Educational Measurement*)

1.	Multiple-choice test items should be based on sound, significant ideas that can be expressed as independently meaningful propositions.
2.	The wording of a multiple-choice item should not follow familiar textbook phraseology so closely that verbal memory without comprehension will provide an adequate basis for response.
3.	Giving the final examination as a pretest will help to identify items that can provide valid measures of specific achievements in the course.
4.	Drafting each multiple-choice item in pencil and double spaced on a separate sheet of paper will facilitate revision of the item and its assembly into the test.
5.	The stem of a multiple-choice item should state or clearly imply a specific direct question.
6.	Item stems including the word »not« and asking in effect for an incorrect answer tend to be superficial in content and confusing to the examinee.
7.	The responses »none of the above« and »all of the above« are appropriate only when the answers given to a question are absolutely correct or incorrect (as in spelling or arithmetic problems).
8.	The intended answer should be clear, concise, correct, and free of clues.
9.	All of the responses to a multiple-choice test item should be parallel in point of view, grammatical in structure and general appearance.
10.	The distracters in a multiple-choice item should be definitely incorrect but plausibly attractive to the uninformed.
11.	The use of compound responses, including an answer plus an explanation, or some combination of two elements, sometimes solves the problem of providing four good alternative answers to a multiple-choice test question.
12.	While most multiple-choice items provide at least four alternative answers, good ones can be written with only two or three alternatives.
13.	A student who selects the correct response to a multiple-choice item by eliminating the incorrect responses demonstrates useful achievement.
14.	To function properly a multiple-choice item must be expressed in carefully chosen

	words and critically edited phrases.
15.	In general the best multiple-choice test items are those that about half the examinees answer correctly.
16.	One can make some multiple-choice items easier by making the question more general and the responses more diverse, or harder by making the question more specific and the responses more similar.
17.	Subsequent, and preferably independent, review of the drafts of multiple-choice test items is likely to improve their quality.
18.	Some of the most effective multiple-choice test questions call for a best answer rather than an absolutely perfect correct answer.
19.	Items testing recall of incidental details of instruction or special organizations of subject matter are ordinarily undesirable.
20.	The stem of a multiple-choice item should be expressed as concisely as possible without sacrificing clarity or omitting essential qualifications.
21.	The responses to a multiple-choice item should be expressed simply enough to make clear the essential differences among them.
22.	True statements that do not provide good answers to the stem question often make good distractors.
23.	The responses to a multiple-choice item should be listed rather than written one after another in a compact paragraph.

Reporting Test Results

Testing may have many purposes. The most appropriate form in which to report results depends partly on the purpose for which they have been collected. In brief reporting raw score marks with no additional information has little value particularly where such marks appear individually. Useful characteristics of performance to consider communicating may include:

- Arithmetic Mean of Marks

This at least provides a reference point against which to judge performance indicated by raw scores. It is unreliable in many ways but better than nothing. It may be considered as providing a test facility index.

- Range of Marks

Though a test may be marked out of 100 only 30% of the mark range may be used, e.g. minimum mark 40, maximum mark 70. This is important if any comparisons or aggregations are subsequently performed.

- Dispersion of Marks

Standard deviation is one indicator of this and is most simply defined as:

The square root of
$$\frac{(\text{the sum of each deviation})^2}{(\text{the number of cases})}$$

i.e. Standard deviation =
$$\sqrt{\frac{\sum (X_1 - X_{av})^2}{n}}$$

The magnitude of this is an indication of the spread of values. Its meaning becomes difficult to interpret where mark distributions diverge considerably from a normal distribution.

- Frequency Distribution of Marks and Percentile Ranks

By calculating the frequency with which marks are obtained and producing a table of cumulative frequency, percentile ranks can be obtained. A candidate's score can then be

reported in a form that relates performance to that of other candidates (see the example given in Table 6).

Score	Frequency	Cumulative Frequency	Cumulative Percentage
10	5	75	100
9	2	70	93
8	12	68	90
7	17	56	75
6	14	39	52
5	10	25	33
4	10	15	20
3	3	5	7
2	0	2	3
1	2	2	3
0	0	0	0
	75		

Thus, a student with a raw score of 5 has a percentile rank of 33, i.e. approximately 2/3 of the students score approximately 1/3 less. Large percentile differences near the 50th percentile or median score may represent small raw score differences and small percentile differences at the top or bottom of the range may represent large raw score differences. Percentile ranks cannot be added; they do provide comparison in terms of a reference group.

- A histogram constructed from the frequency tabulation of raw scores provides a useful diagrammatic representation of several test characteristics and is a valuable piece of information to provide. Its shape is a useful guide to the meaning of such statistics as the standard deviation etc.

- Profile Reporting

Where a test has several components which are intended to measure different types of performance, profile reporting may be appropriate. This involves reporting separately on different areas of skill or knowledge rather than reporting an aggregate mark derived from several Science tests which are directed towards measuring different characteristics. Aggregation may well degrade the quality of the information that can be provided if results in different areas do not inter-correlate. For example suppose we have the results of three Science tests:

Candidate	Paper I Recall Information	of	Paper II Comprehension/Application	Paper III Practical Skills	Aggregate Mark
Kwasi	52		61	38	151
Kofi	37		41	72	150
Hamid	61		49	31	141
Nasreen	70		71	48	189
David	30		28	62	120

At a glance we can see that Paper III results do not appear to correlate positively with Paper I and II results. If we produce a single aggregate mark by addition we disguise useful information. Note that although Kwasi and Kofi have similar aggregates Kwasi has achieved his mainly through good performance on Papers I and II; Kofi, on the other hand, achieved his largely through a high score on Paper III. This clearly has implications for any individual attention that may be given to them to improve their performance as well as whether it is intended to allow performance in one area to compensate for that in others.

One way of reporting results like those above is to draw up a matrix of categories to provide indications of performance on different test types, e.g.

Name Kofi		Performance		Subject Science	
Recall of Information			X		
Comprehension/Application			X		
Practical Skills				X	
	Well below average	Below average	Average	Above average	Well above average

This procedure is not perhaps valuable when marks between tests are highly correlated. However, high correlation between test marks on different tests designed to measure different capabilities does raise questions concerning whether they do in fact measure different things. A scattergram of marks obtained on two tests (or different parts of the same test) provides a simple diagrammatic indication of the extent of correlation between measures of performance on different traits. Several statistical methods for calculating such correlations are available. See for example:

Standardizing Test Scores

Simple addition of raw scores to produce aggregates for overall ranking can be very misleading. Table 8 and 9 show this.

	Subject								Total
	A	B	C	D	E	F	G	H	
1. Kwasi	80	50	80	50	25	60	32	42	419
2. Kofi	95	70	15	80	40	40	30	48	418
3. Hamid	30	80	65	20	70	55	34	40	414
4. Nasreen	81	60	30	60	60	36	44	42	413
5. David	94	20	75	50	50	30	47	46	412
6. Kim	60	90	20	75	30	45	45	46	411
7. Naisun	100	40	40	70	20	40	52	48	410
8. Cream	10	75	50	55	30	60	56	45	381
9. Sam	31	30	60	70	48	35	55	50	379
10. Nona	80	10	50	25	30	70	50	55	370
11. Sharifah	20	80	50	30	40	55	60	52	367
12. Vina	70	25	10	50	40	50	55	60	360

1. Vina	67	19	0	50	40	50	83	100	409
2. Sharifah	11	88	29	17	40	63	100	60	408
3. Nana	78	0	57	8	20	100	67	75	405
4. Sam	23	25	71	83	56	13	83	50	404
5. Cream	0	81	57	58	20	75	87	25	403
6. Naisun	100	38	43	83	0	25	73	40	402
7. Kim	56	100	14	92	20	38	50	30	400
8. David	93	13	93	50	60	0	57	30	396
9. Nasreen	79	63	29	67	80	15	47	10	390
10. Hamid	44	88	79	0	100	63	13	0	387
11. Kofi	94	75	7	100	40	25	0	40	381
12. Kwasi	78	50	100	50	10	75	7	10	380

Table 9 indicates the results of applying a simple scaling correction for the range of marks used. Rank order is completely reversed.

$$\text{Scaled Score} = (\text{Raw Score} - \text{Minimum Raw Score}) \times \frac{100}{\text{Range of Raw Scores}}$$

This correction does not standardize the marks and give equal weightage to each component though it is a better reflection of relative performance than raw scores alone.

Standardization of marks on different tests to give equal weightage to all components is most simply accomplished by computing a corrected score that relates performance on each paper to the mean and standard deviation of marks obtained. A standard score or Z-Score can be simply derived

$$\begin{aligned} \text{Z - Score} &= \frac{\text{Raw Score} - \text{Mean Score}}{\text{Standard Deviation}} \\ &= \frac{\text{Deviation of Raw Score}}{\text{Standard Deviation}} \end{aligned}$$

This can be converted to a mark that does not go negative by employing an arbitrary scale with, for example, a mean of 50 marks and a standard deviation of 15.

The standardized mark then becomes

$$(\text{Z-Score} \times 15) + 50$$

Summing standardized marks of this kind does give equal weightage to each component score in terms of relative performance under certain conditions.

Suppose we have 5 candidates taking 3 papers and we want an indication of overall performance, weighting each paper equally. The aggregate and standardized aggregate scores are as shown below.

Candidate	Paper 1	Paper 2	Paper 3	Aggregate	Rank	Standardized Aggregate	Rank
Kwasi	68	48	63	179	1	164	1
Kofi	60	74	43	177	2	151	3
Hamid	46	59	71	176	3	157	2
Nasreen	40	87	37	164	4	131	5
Kim	76	32	51	159	5	148	4
Mean Raw Score	58	60	53				
Standard Deviation	13.39	19.26	12.52				

Under certain conditions standardization can change the rank order of candidates when aggregate scores are compiled. It will almost certainly modify the magnitude of differences in relative performance. Standardization techniques of this kind are only really appropriate to large populations where distributions approximate a normal curve. They can themselves be misleading when applied to very skewed distributions.

2.3. Developing and Administering Tests in Primary Science and Agriculture - A Practical Exercise

Udo Bude & Keith Lewin

This chapter discusses the need to improve skills in the creation of assessment items used in primary school leaving examinations in Africa. The first section provides a rationale for the

importance of this kind of activity. The second gives an overview of a workshop process that we have found useful in encouraging reflection and developing skills. It includes examples of workshop tasks which provide a condensed experience of the test construction, review, administration, analysis and revision process based on material used in the workshop held in Nyeri in Kenya in 1994. The test instrument created was piloted in Kenyan schools. Results were analysed using simple techniques to provide a basis for appraisal of performance characteristics and quality of the items that had been generated.

The need to improve the quality of test and examination questions is well established. Studies in Africa from the early 1980s and before attest to great variations in the technical attributes of test items between countries, the importance of examination content in influencing teaching and learning, and the possibility that existing procedures and instruments may lack adequate reliability, validity and curricula relevance. (ILO/JASPA 1981; Somerset 1982; Oxenham (ed.) 1984; Somerset 1988; Kellaghan and Greaney 1992). In this section we focus especially on the construction, analysis, and review process for multiple-choice styles of test construction. This is not because they are the only or perhaps the best way of assessing achievement at the end of the primary cycle. It is because they are indeed the most widely used style of test construction at this level. They offer advantages of cost, convenient administration, wide sampling of the curriculum, reliable marking, and convenient statistical methods of checking for some types of fraudulent behaviour. The well-known limitations include the inability to assess some types of valued outcomes, uncertainties over how correct responses are generated, influences on the curriculum in action that may discourage some types of classroom activity, and the risk that school achievement is narrowly defined by performance on them. The advantages can only be realised where technical quality is high, administration is efficient and honest, and item writers can set questions which span an appropriate range of cognitive outcomes. The disadvantages are also limited where these conditions are satisfied. This is why adequate training and support for those who become item writers is needed.

Many of those who become test constructors have backgrounds and experience that only prepare them partially for the demanding roles that they are required to play. They may have taught for many years in the school system at a particular level and find themselves required to examine at other levels (especially the case with primary school leaving examinations which are often constructed by those whose teaching experience is at secondary level). They may have worked in different parts of the Ministry of Education unrelated directly to examining and test construction as administrators, or inspectors. Even those with a background in curriculum development are unlikely to have a significant amount of specialised training in examination construction. A wide range of skills are needed for effective test construction. These include a sophisticated awareness of children's thinking in Science at different ages and the ability to translate this into questions which have appropriate levels of cognitive demand and are expressed using language structures that are accessible to children (especially the case where the medium of instruction is not the mother tongue). It is also important to have detailed knowledge and understanding of the curriculum and its intended learning outcomes, preferably based on the experience of teaching the relevant curriculum at an appropriate level. Statistical competence is essential if quantitative data on performance is to be interpreted meaningfully and pilot studies used to improve item performance against explicit criteria. Skill in deciding what balances should be struck between the demands on examinations to provide valid and reliable data for selection, to certify competencies against known standards, and to use the power of examinations to influence learning and teaching and support the achievement of curricula goals.

African examination authorities have become more and more aware of the need for systematic staff development to develop necessary skills and in many cases donors have assisted in supporting training programmes for staff. Though the picture is changing it is still all too frequently the case that the construction falls short of the standard that pupils and parents have a right to expect and governments need to have confidence in results that determine life chances and underpin quality improvement in education and increased levels of achievement. Any strategy that attempts to improve matters must depend on a mixture of in-house training and support, mutual assistance between examining authorities across Africa which can share expertise and training opportunities and access to external knowledge and expertise which reflects the most recent developments in the field.

The material presented in the rest of this chapter provides a basis for training and development exercises that can be mounted at within countries and at regional level and can readily be adapted to reflect specific demands of particular national systems. Thus, though we have chosen to base this exercise on the Kenya Std. 7 syllabus for Science and Agriculture, this was only chosen for reasons of expediency given that the practical application of the trial test materials was in Kenyan schools.

The Overview which follows explains step by step common procedures followed in the writing of items which can be used to develop a trial test paper of any number of items. It explains the process through which the pilot instrument was tried out in six schools, the results statistically and judgementally analysed, and the process of critical review that followed. In reality a revised test instrument would then have been designed and re-piloted to improve the quality of the test.

The item writing task is made up of a number of sections denoted **Information** and **Assignments**. Short summary contextual information is provided in the Information briefs which prepares the ground for the Assignments, each of which defines a task to be undertaken in groups. The trial test paper that was produced is then reprinted in full. Each group administered the test according to an agreed procedure that arose from discussion of good practice. This included standardised administration conditions, subsequent interviews with relevant teachers which were written up commenting on opinion concerning the quality and relevance of the items in the test instrument, and reflections on the administration of the test in relation to ideal conditions. Test data was then analysed manually by the groups. This was deliberate to ensure deeper understanding of the process than would be the case if results were simply entered into a computer to generate item statistics that might not be understood at all. The groups were then invited to reach judgements on the performance of the items based on the data generated. This was particularly fruitful when outcomes were compared with anticipated characteristics of performance challenging groups to explain why patterns emerged that they had not expected (thus some items were easier than anticipated and some more difficult, some differentiated more strongly between boys and girls than others in ways which were not predicted).

The tasks we have devised are designed to make use of the previous experience of participants. We deliberately tried to include in the groups individuals from an examination department background and those who had experience of curriculum development. It was also useful to spread subject expertise in the groups across the life and physical sciences.

In summary the process outlined below consists of:

- Basic Information Sheets containing information relevant to each task (Overview, Info 1)
- Assignment sheets specifying group activities used in construction, administering and analysing tests (Assignment 1, Infos 2-4)
- Test item development procedures (Assignments 2-4, Info 5)
- Item prediction (Assignment 5)
- Trialling in schools (Assignment 6, Test Administration)
- Marking procedures (see Marking Scheme for the Test Paper)
- Analysis of Items (Assignment 7, Infos 6 and 7)

At the end of this chapter some conclusions are reached on the further development of tests and enhancements to the workshop process.

2.3.1. The Item Writing Task

OVERVIEW

The task is to create a 30 item test which will be set in six local schools. The work will be organized in three working groups. Each group will produce ten to fifteen items. The specification for the items is as follows:

1. Items should relate to the following Topic Areas in Grade 7:

Science	Agricultural Science
Living Things Properties of Matter Weather Making Work Easier	Soil Conservation Diseases and Pests Simple Farm Accounts

Note: Read Info 1: Primary Education Syllabus, Kenya, May 1992

2. All items should be above the level of knowledge recall and recognition, i.e. at least comprehension or application level.

3. Each group should produce a total of 10-15 items, at least two of each category which assess learning outcomes related to:

- Observation - skills of looking and seeing
- Classification - skills of categorising cases according to rules
- Measurement - skills of deciding how or what to measure
- Recording Data - skills of deciding how to record data
- Interpretation of Data - skills of interpreting data e.g. graphically

4. The procedure will be as follows:

- all individuals draft at least two items after agreeing with group leader which areas they will cover to avoid unnecessary duplication;
- each individual discusses each item written with other members of the group and revises the items;
- group and group leader review all items produced to arrive at a list of about 10 items covering the curriculum areas identified above in Science and Agricultural Science;
- each group constructively reviews work of other groups to improve items;
- final list of items constructed;
- 30 item instrument applied to pupils in six schools;
- test papers marked by each group for the pupils tested (i.e. each group marks two schools only);
- facility values and discrimination index calculated for each item;
- each group reviews results on technical and other criteria;
- summary report written commenting on items.

5. Test administration will take place in six schools - two high achieving, two average, two low achieving. Each group will be responsible for administering the test in two schools.

6. Each group will split into two sub-groups - one for each school visited. One member will make sure the items are administered effectively and answer any questions about the test. Other group members should prepare a semi-structured interview schedule to ask about assessment issues e.g. items on last years Primary School Examination e.g. which type of item pupils found most difficult/easy? Why? Which were »unfair«? Which did girls do best/worst? How do teachers prepare pupils for the examination? The visit should include reviewing pupils' work books and establishing which text books are available and how they are used. If possible, examples of test questions used by teachers should be collected.

7. Group leaders will organize the marking and simple statistical analysis and take responsibility for coordinating the production of a short written report.

8. Proposed Timetable

→ DAY 1	
am	Task introduction/
am	Writing individually
pm	Peer critiques/
pm	Item revision
→ DAY 2	
am	Group critiques of items/Revision
pm	Draft selection and compilation
→ DAY 3	
am	Inter group item critiques/Final selection
am	Interview preparation
pm	Production and copying/
pm	Predictions of item performance
→ DAY 4	
am	Administration and interviewing in schools
pm	Discuss and write up interview notes
→ DAY 5	
am/	Marking and statistical analysis
pm	
→ DAY 6	
am/	Analyses and report writing
pm	
→ DAY 7	
am	Reporting/Discussion of findings

*Info 1: Primary Education Syllabus Kenya - May 1992 (abstracts)**

* from: Kenya Institute of Education (May 1992) Primary Education Syllabus. Vol. I. Ministry of Education. Republic of Kenya.

SCIENCE (abstracts)

STANDARDS IV-VIII

I. INTRODUCTION

The work contained in this syllabus is presented in form of units. The units for each class need not be dealt with in the order in which they are written. The teacher should feel free to tackle any unit as might be dictated by environmental conditions and/or children's interests.

The teacher should, at all times, realise that SCIENCE IS DOING not just being told and therefore pupils should be actively involved in learning.

The teacher should use a variety of teaching methods, for example, nature walk, group activity, project work, demonstration, and so on, in handling the content.

Pupils and teachers should collect most of the resource materials required. Other materials should be constructed by pupils under the guidance of the teacher.

II. AIMS OF SCIENCE TEACHING

The main aims teaching Science in primary schools are:

1. To enable children to acquire and preserve certain useful ATTITUDES about themselves and their relationship with the environment.
2. To enable children to acquire a basic scientific KNOWLEDGE.
3. To enable children to acquire certain manual and thinking SKILLS which are useful in solving practical problems.

III. GENERAL AND SPECIFIC OBJECTIVES

At the end of the course the learners should be able to:-

1. Apply a problem solving approach to all investigations.

Specifically the learners should be able to:

- (a) Identify problems, design investigations, examine evidence and take responsibility for their own learning.
- (b) Demonstrate the skills of observation classification, measurement, recording, communication, making predictions, formulating hypotheses from evidence.
- (c) Analyse cause and effect relationships, control variables and draw rational conclusions.
- (d) Recognise the relevant information required to solve a particular problem.
- (e) Use a variety of sources for acquiring any necessary information.
- (f) Put in order the information gained during their study of science.
- (g) Seek new ways of finding further information.
- (h) Demonstrate feelings of self-confidence and a sense of working together.

2. Identify the major factors of, and develop and use appropriate skills and technologies for solving the problems relating to:-

- (a) Conservation and utilisation of energy and other resources.

Specifically the learners should be better able to:-

- (I) Observe and record how plant, animal, and human resources are used in their community.
- (II) Devise other methods of utilising the above resources.
- (III) Observe and record various sources and uses of energy.
- (IV) Observe and analyse methods of conserving energy in the community.

(V) Make simple machines that make work easier.

(b) Use of communication.

Specifically the learners should be able to:-

Communicate using a variety of techniques e.g. writing, drawing, modelling, graphing, sketching and making charts, oral (debates, interview, discussions, drama), listening and interpretation of information.

(c) Public health and hygiene in the community.

Specifically the learners should be able to:-

(I) Practise care when handling medicines.

(II) Discuss dangers of drug abuse.

(III) Prevent drug abuse by educating self and others.

3. The learner should be able to collect, record, interpret and communicate for rational decision making.

This objective has no specific objectives. Rather it will be achieved through work in all other objectives.

4. The learners should be able to develop flexibility and adaptability in solving problems.

Specifically the learners should be able to:-

(a) Develop a willingness to listen.

(b) Develop a feeling of concern, care and consideration for others.

(c) Accept new ideas.

(d) Experiment with new ideas.

(e) Co-operate in group activities (family, class and community).

(f) Accept and participate in desirable cultural activities and practices.

(g) Share new ideas in the family, class and community.

(h) Educate family and community on these values and importance of new ideas.

(i) Adjust to new changes in the environment,

(j) Demonstrate honesty, patience and accuracy in solving problems,

(k) Recognise the need for flexible planning.

5. The learners should be able to adopt solutions to problems of management and conservation of available resources.

Specifically the learners should be able to:-

(a) Identify various human resources available in the community.

- (b) Relate use of resources for meeting family needs to values, goals, standards and available resources.
- (c) Consult with local experts when solving problems.
- (d) Plan and use labour saving devices to save time and energy.
- (e) Analyse and encourage recreational activities conducive to good physical and mental health.
- (f) Make economic contribution to the individual and the community.

6. The learners will be better able to promote, preserve and evolve their national heritage for their cultural, spiritual and economic development.

Specifically the learners should be able to:-

- (a) Make excursions to places of national cultural and scientific interest.
- (b) Use their required scientific skills and knowledge to re-design and improve traditional tools and implements.
- (c) Investigate the scientific validity of traditional beliefs and practices.
- (d) Acquire a balanced perspective on such beliefs and practices.
- (e) Take independent decisions when confronted with conflicting viewpoints arising from such beliefs and practices.
- (f) Discourage those beliefs and practices which are detrimental to community health, spiritual well being and economic development.

7. The learners will be better able to identify and utilise opportunities for productive work in the home and the community.

IV. SUMMARY OF CONTENT

- Standard Four:** Weather and Astronomy, Living Things, Heat, Making Work Easier, Properties of Matter, Health Education, Soil.
- Standard Five:** Living Things, Soil Balancing and Weighing, Heat, Sound, Properties of Matter, Health Education.
- Standard Six:** Weather, Living Things, Environment, Properties of Matter, Electricity, Making Work Easier, Health Education.
- Standard Seven:** Weather and Astronomy, Living Things, Environment, Soil, Properties of Matter, Electricity, Making Work Easier, Health Education.
- Standard Eight:** Living Things, Environment, Properties of Matter, Energy, Light, Making Work Easier, Health Education.

STANDARD SEVEN

Unit 1: Weather

(I) Constructing weather measuring instruments

- rain gauge
- wind vane
- wind sock

- liquid and air thermometer

Unit 2: Living Things

(I) Inter-dependence between:

- plants
- animals
- plants and animals

(II) Blood circulatory system

- heart
- blood vessels
- blood

(III) Importance of blood circulation

(IV) Composition of blood

- plasma
- blood cells
- platelets

Unit 3: Environment

(I) State the meaning of environment

(II) Describe components of the environment

(III) Pollution in:

- soil
- water
- air

(IV) Conservation of:

- soil
- water
- air
- plants
- animals

Unit 4: Properties of Matter

(I) States of matter (atoms and molecules should NOT be mentioned)

(II) Changes of state e.g. water to vapour

(III) Comparing weights of equal volumes in different substances

(IV) Density: Activities to demonstrate

- Units of density (No calculation of density required)

(V) Composition of air

Unit 5: Electricity

(I) Electric circuits in series and parallel using

- batteries and bulb (Ohm's Law not required)

Unit 6: Making Work Easier

(I) Motion

- making things move
- stopping moving objects

(II) Describe what force is

- mention units of force (No calculations required)

(III) Describe what friction is

- state advantages and disadvantages of friction
- reducing friction

(IV) Making work easier using levers e.g. wheel barrow

(V) Identifying positions of:

- fulcrum
- load
- effort

Unit 7: Health Education

Drug Abuse

(I) The health, social and economic implications of abusing the following drugs:

- tobacco
- alcohol
- miraa (khat)

(II) Demonstrate the effect of cigarette smoke on a piece of wet white material.

AGRICULTURE (abstracts)

I. INTRODUCTION

This course is designed for upper primary classes starting at Std. IV. The pupils are expected to have gone through the integrated Lower Primary Science Course as a foundation upon which this course shall be built.

By Standard IV, it is expected that the pupils will have developed to a stage where they can be introduced to practical Agriculture.

The course is designed to help both the pupils who will end their formal education at K.C.P.E. level and also for those who will go for further education. For this purpose, the course is meant to give pupils a foundation for real life agriculture if they take farming after school and also for further agricultural education in post-primary institutions.

The following are proposed guidelines which will help the teacher to organize and teach Agriculture.

1. Scheming

The teacher who will handle this subject should be familiar with local environmental changes and when they occur. This is important so that projects may be well timed to concur with these changes. Teachers also should as much as possible see that they relate the school activities with those being carried out in the community e.g. weeding - when the local community is weeding their crops....

2. School Shamba

In many instances, pupils' interest towards Agriculture is killed through use of school garden as a place of punishment. The garden should be looked upon as a place of learning, a laboratory to carry out experiments and a resource for ideas.

II. OBJECTIVES OF PRIMARY SCHOOL AGRICULTURE

1. To demonstrate through practical experience that agriculture is a profitable and honourable occupation.
2. To create awareness of the importance of agriculture in the daily life of various communities and Kenya as a whole.
3. To assist the pupils practically acquire agricultural knowledge and skills which are relevant and useful to their lives.
4. To stimulate genuine interest and develop positive attitudes leading towards active participation in agriculture.
5. To ensure that the schools take an active part in rural development by integrating agricultural activities in the school curriculum.
6. To develop self reliance, resourcefulness, problem solving ability and occupational outlook in agriculture.

STANDARD SEVEN

Unit 1: Diseases and Pests

(a) Crop Pest and Diseases

- (I) Observing and identifying the differences between healthy and unhealthy crops.
- (II) Identifying common symptoms of various diseases and pests on crops.
- (III) Identifying nutrient deficiency in crops.
- (IV) Finding information on and practising methods of control and treatment of the common diseases in crops.
- (V) Methods of maintaining soil fertility.

(b) Livestock Diseases and Parasites

- (I) Observing and identifying the differences between healthy and unhealthy animals.
- (II) Identifying symptoms of various diseases and parasites on animals.
- (III) Finding information and practising methods of control and treatment of the various diseases and parasites in livestock.

Unit 2: Animal Feeding

- (I) Discussing the constituents of balanced diet e.g. carbohydrates, proteins, vitamins, mineral salts, water.
- (II) Identifying common feeds on the farm for feeding livestock.
- (III) Identifying the nutritional deficiency symptoms in livestock.
- (IV) Practising feeding the animals on school/home farm with balanced diet.

Unit 3: Soil Conservation

- (I) Discussion on soil conservation methods.
- (II) Visiting farms and identifying soil conservation methods used in the community.
- (III) Practising some of the methods of soil conservation on school/home farm or garden e.g. mulching, cover cropping, terracing, contour farming, tree planting.

Unit 4: Fish Farming

- (I) Finding information on the following:
 - (a) siting a fish pond
 - (b) construction of fish pond
 - (c) type of fish kept
 - (d) fish rearing, fish harvesting, fish preservation, fish marketing.
- (II) Constructing a fish pond where possible.
- (III) Rearing some fish in the pond.
- (IV) Visiting some fish farms where possible.

Unit 5: Storage

- (I) Studying the various methods of storage in the community both traditional and modern by visiting:
 - (a) large/small scale farms
 - (b) dairies
 - (c) marketing boards' stores
 - (d) co-operatives stores
- (II) Discussing storage of farm produce, agro-chemicals.
- (III) Observing and practising the methods used in keeping the stores clean, safe, dry and free from pests.

Unit 6: Youth Organizations

- (I) Discussing the roles of youth agricultural organizations e.g. Young farmers Clubs, 4-K clubs (Kuungana, Kufanya, Kusaidia Kenya).
- (II) The relationship of Y.F.C. to the Agricultural Society of Kenya.
- (III) Identifying the roles of youths in these organizations at school, home and national level.

(IV) Participating in community projects which are agriculturally oriented e.g. construction of dips, water projects, freedom from hunger walks, national youth weeks etc.

(V) 4-K Club activities.

(VI) Participating in activities of agricultural organizations e.g. agricultural shows, farmers field days, etc.

Unit 7: Methods of Grazing

Discuss the following methods of grazing including their advantages and disadvantages.

(I) Rotational grazing

- (a) paddocking
- (b) strip grazing
- (c) tethering

(II) Zero grazing (stall feeding)

(III) Herding

NB: Visiting local farms to see how these methods are carried out.

Unit 8: Simple Farm Accounts

(I) Discussing and identifying documents used in farm accounts.

(II) Importance of farm accounts.

(III) Discussing and identifying types of farm accounts.

Assignment 1: Basic Instructions

- 3 working groups - each group produces 10-15 items.
- All items should be above the level of recall and recognition.
- Items to cover topics and skills identified in table of specification.
- Each person drafts two multiple-choice items for different learning outcomes.
 - Items should have a key and three distractors.
 - Items should fit into half a page or less.
 - Art work should be sketched.
- Items should be reviewed by half groups and revised.
- Group leaders should coordinate production of items.
- Each item should indicate the name of the developer and the topic/skill areas intended.

REMEMBER THE ART WORK

IF STUCK LOOK AT RESOURCE MATERIALS AND DISCUSS WITH ANOTHER GROUP MEMBER/FACILITATOR!

EXAMPLES:

Table of Specification - Topics by Skills					
Content	Observation	Classification	Measurement	Recording	Interpretation
Living Things					
Properties of Matter					
Weather					
Making Work Easier					
Soil Conservation					
Diseases and Pests					
Simple Farm Accounts					

GROUP 1		
Name	Topic	Skill
1. Mwanza	Living Things	Observation
	Weather	Classification
2. Mariam	Soil Conservation	Measurement
	Properties of Matter	Interpretation
3. Grace		
4. Moshoeshoe		

Info 2: Advantages and Disadvantages of Multiple-Choice Questions

Multiple-choice questions require pupils to select the correct answer from among several options - typically options A through D or True and False. In other types of questions pupils construct their responses, in writing, speaking, or perhaps by creating a piece of art or music.

Multiple-choice question types can have a dramatic effect on what happens in the classroom. If the examination primarily contains items which have pupils selecting from among options, then pupils are likely to spend much of their time preparing for the test with worksheets in which they select the correct answer. This type of behaviour is less likely to help pupils make the extensive web of mental connections which help them to understand and use what they learn in school. However, multiple-choice questions are a very efficient way to measure a large body of knowledge and are likely to be an integral part of any test developed and used for large-scale testing.

Advantages	Disadvantages
<ul style="list-style-type: none"> • Because multiple-choice questions take little time to answer, a test can measure a broader range of content than is possible with a test which relies solely on essay items or performance tasks. • They are less costly to score. If specially prepared forms are used, they can be machine scored, allowing thousands of answer sheets to be scored in a very short period of time. 	<ul style="list-style-type: none"> • It is more difficult to design multiple-choice questions which measure higher levels of thinking and problem solving. • It is more difficult to design questions which measure more complex, real-life types of skills and thinking. • They take more time to develop because of the need to construct four or five response choices. • Multiple-choice tests promote multiple-choice teaching -that is, teaching where students are always looking for the one right answer.

Advantages	Disadvantages
<ul style="list-style-type: none"> • They are an efficient way to measure recall of factual knowledge and some skills. 	<ul style="list-style-type: none"> • There is a high chance of being able to get the correct answer by guessing, which is not the case with performance tasks or essays. If a multiple-choice question has four options, the student has a 25 percent chance of guessing the item correctly.

Note: This information is based on Capper, Joanne (March 1994) Testing to Learn... Learning to Test. A Policymaker's Guide to Better Educational Testing. Executive Summary. Academy for Educational Development, Washington D.C., p. 23.

Info 3: Format of Multiple-Choice Items

STEM	Pupils of Makumbe Primary School have observed for a number of years that their maize crops get damaged by strong winds.		
	What should they do to stop the damage in future?		
	A. Apply more fertilizers to their maize crop.	X	
DISTRACTOR	B. Grow a short maize variety.	X	OPTIONS (A-D)
	C. Intercrop maize with beans.	X	
KEY	D. Plant trees around their maize plot.	X	

NOTE:

- The stem should not test trivia.
- The stem should set an unambiguous task.
- The stem should not contain any kind of »cues« (hints, signals etc.).
- The stem and all options should be checked for reading level (especially if language of instruction differs from regional/local languages).
- The options should all be plausible, comparable and similar in essential particulars.
- Avoid grammatical clues in the options.
- The key must be absolutely correct.
- The distractors must be indisputably wrong.

Info 4: Suggestions to Consider in Writing Higher-Order Items

- Avoid questions that start with words such as who? what? when? where? These always solicit factual information.
- Use words like why?, because... These require students to reason.
- Give a set of conditions and ask the students to predict a future result.
- State a problem and ask the students to suggest a solution, e.g. separating salt from sand.
- Use as much stimulus material as possible, e.g. diagrams, pictures, graphs, tables. These are easier to set questions on than long winded statements.
- While thinking of the item to write, think in terms of problems to be solved rather than facts to be remembered.

- Always strive to use real life problems instead of abstract concepts.

PRACTICE MAKES PERFECT. KEEP ON TRYING. DO NOT GIVE UP

Assignment 2: Item Writing and Gender Aspects

Recent Curriculum Revisions/Reforms have introduced changes designed to make subjects like Science and/or Agriculture more relevant. An important aspect of this is to ensure that new courses are equally relevant to boys and girls. Items developed for examination need to be checked thoroughly for gender bias.

As the next step review the items developed so far in your working group and consider the following questions:

- Does the language used for the items show gender bias? (Is »he« used more than »she«, boys' names more than girls' names, first person or third person accounts of situations etc.).
- Do the items cover situations and topics likely to be of interest to both girls and boys and circumstances they have experience of? (Work environments, home life, school experiences).
- How many women/girls - men/boys feature in the items?
- Do illustrations show women/girls in a positive or negative way?
- Are there any differences between female/male illustrations? (Who is shown in a modern or in a traditional context. Who features in what kind of situations?).
- On which questions are boys likely to perform better; on which are girls?
- Do any items relate specifically to concerns likely to be of special interest and relevance to girls or boys? (infant care, child birth, inheritance, farming activities etc.).

Info 5: Gender Stereotypes in Science Teaching/Learning

Girls appear to underachieve in most scientific areas in comparison to boys in many countries in Africa, especially when their participation rates are taken into account.

This cannot be explained by biological or intellectual factors, but rather by peer pressure and cultural expectations which influence the design of school curricula, the nature of examinations, which may include items which favour boys rather than girls, and the way Science is taught in schools (some topics may be less interesting for girls; teachers reinforce stereotypes prevailing in society; boys receive more attention; textbooks contain gender bias etc.).

Working against Stereotypes:

- Given appropriate conditions girls achieve in Primary and Secondary Science as well as, if not better, than boys.
- Research has produced no reliable evidence that girls have less Science ability than boys have.
- When girls see that Science addresses their concerns and values they show just as much interest as boys do.
- When girls recognise that a problem relates to areas that are important to them, and they

are free to include contextual aspects, they are able to produce valuable and realistic solutions to problems.

- Though differences may exist between boys and girls (e.g. in spatial awareness, verbal reasoning, numerical facility) they are small and may differ between populations, and are unlikely to have significant effects on school achievement in Science. Given time, girls can weigh up aspects of the complex problems they perceive and come to conclusions which they can justify.
- Girls have the same rights as boys to acquire relevant skills and understanding to serve as a base for citizenship and employment.

Adapted from: Harding, Jan (July 1992) Breaking the barrier. Girls in science education.

FOR FURTHER READING:

Eshiwani, G. (1988) Participation of Girls in Science and Technology Education in Kenya. Ann Arbor. Michigan State University (Working Paper 168).

Obura, Anna P. (1991) Changing Images. Portrayal of Girls and Women in Kenyan Textbooks. Nairobi/Kenya.

Tsayang, G.T. & Ngwako, A.D. et al. (eds.) (Sept. 1989) Gender and Education. Proceedings of a Workshop. Occasional Paper No. 2, University of Botswana. Gaborone.

Assignment 3: Individual Review of Test Items and Prediction of Outcomes

- In terms of difficulty (D)
- Discrimination (do the best students get the item correct?)
- Most powerful distractor
- Differences between boys and girls

Make a table of the following kind:

Item Number	Difficulty	Discrimination	Distractor	Gender Difference
1	H	L	A	B
2	M	M	C	G
3	L	H	B	B
4	M	L	A	G

H = High, M = Medium, L = Low
 B = Boys>Girls G = Girls>Boys

Assignment 4: Further Review of the Draft Test

- Split into three groups.
- Analyse test items in sub-groups of two or three. Participants identifying key and the length of time to complete. Do not stop to argue. Note items where there is a problem.
- Meet in groups 1, 2 and 3. Return to problem items and suggest essential editing. Indicate items to be discarded. Write down suggested edits and pass on for correction. Each group concentrates on parts of the test:

For example:

- Group 1 to concentrate on items 1-14.
- Group 2 to concentrate on items 15-26.
- Group 3 to concentrate on items 27-38.

Make suggestions on other items if more were produced and if there is time.

- Check clarity of stem - can you understand what you are asked to do?
- Is key clearly correct?
- Are distractors suitably balanced?
- Is diagram ok?
- Is language ok? (Consider primary students' level of English).
- Select the final 30 items for the **Test Paper**.

TRIAL TEST PAPER: Science and Agriculture (incl. Making Scheme)

Time: 1 hour

1. This booklet consists of 30 questions.
2. Write your name and the name of your school below

your name _____
name of your school _____

3. Indicate below by a tick (✓) whether you are a boy or a girl

boy ()
girl ()

4. For each of the questions in this booklet four answers are given. The answers are lettered A, B, C, D. In each case only one of the answers is *correct*. Choose the correct answer and circle it using a pencil as shown in the following example:

Example

The process by which water is lost to the atmosphere from the soil is called

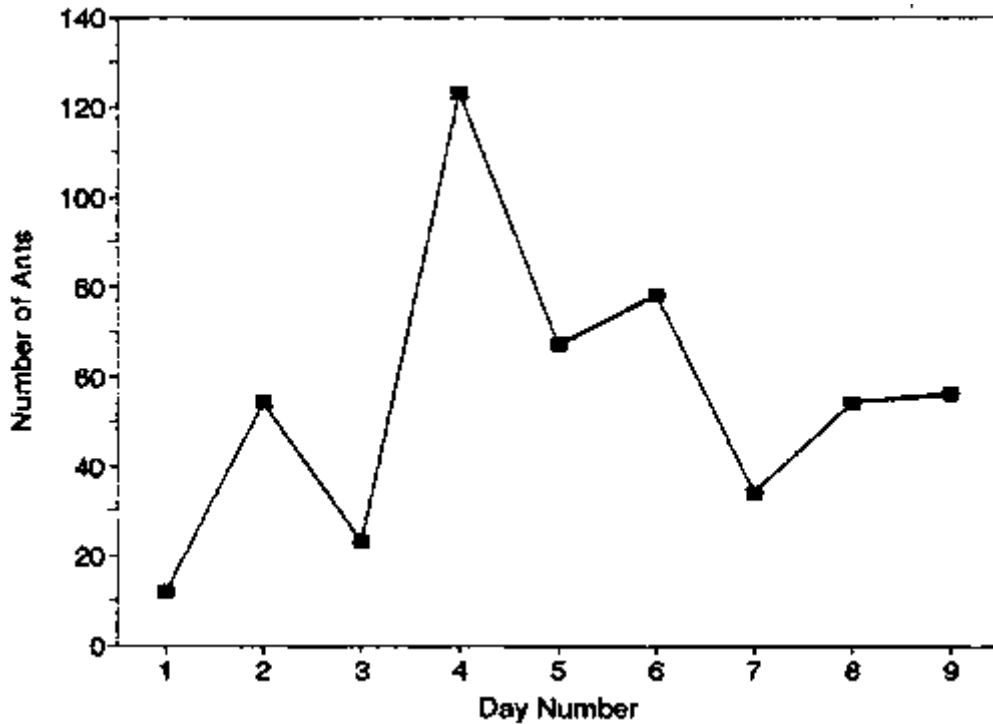
- A. Condensation
- B. Evaporation**
- C. Distillation
- D. Transpiration

The correct answer B is circled, if you want to change the answer that you have circled rub it out and circle a new answer.

5. Answer ALL the questions

Note: This TRIAL TEST PAPER was developed by participants from Eastern and Southern Africa during a workshop on »Writing of Test Items for Primary School Leaving Examinations«, Nyeri/Kenya 1994, jointly organized by the German Foundation for International Development (DSE) and the Kenya Institute of Education (KIE).

1. Standard 7 pupils at Mathaithi Primary School studied and counted ants on a 4 square metre piece of land. They kept the following record for 9 days.



Number of Ants on Different Days

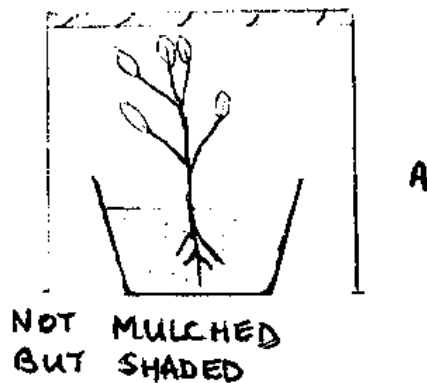
On which days did they find less than 50 ants?

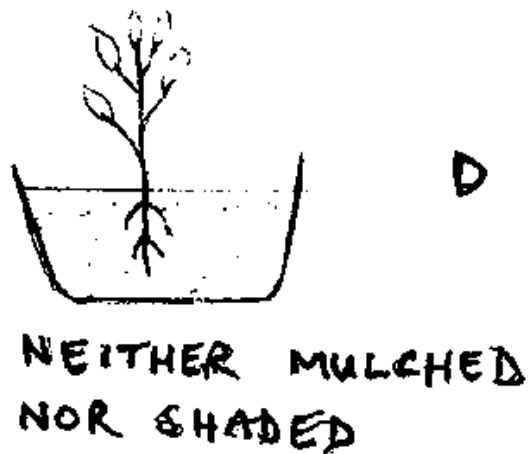
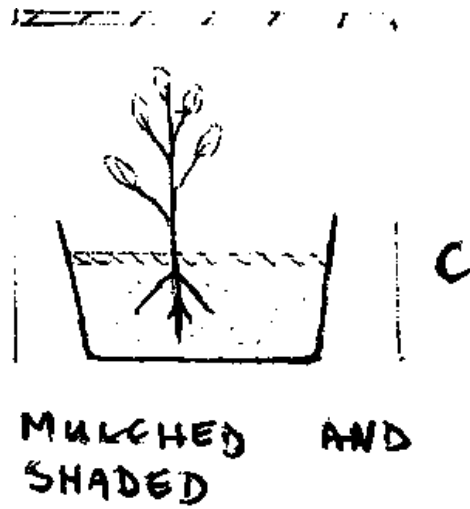
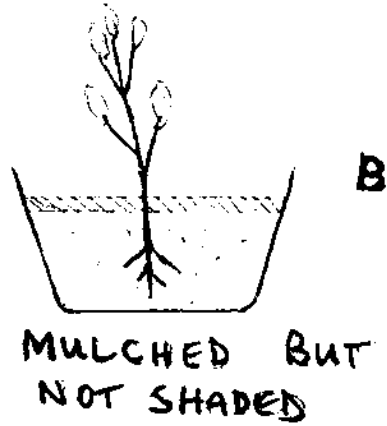
- A. Days 1, 3 & 6
- B. Days 3, 5 & 6
- C. Days 1, 3 & 7
- D. Days 2, 4 & 5

2. Which one of the following weather instruments is INCORRECTLY matched with what it is used to measure?

<u>Instrument</u>	<u>Use</u>
A. Anemometer	Windspeed
B. Thermometer	Temperature
C. Hygrometer	Rainfall
D. Barometer	Pressure

3. John planted some seedlings in pots as shown below.





Which of the pots will conserve more water

- A. Not mulched but shaded
- B. Mulched but not shaded
- C. Mulched and shaded
- D. Neither mulched nor shaded

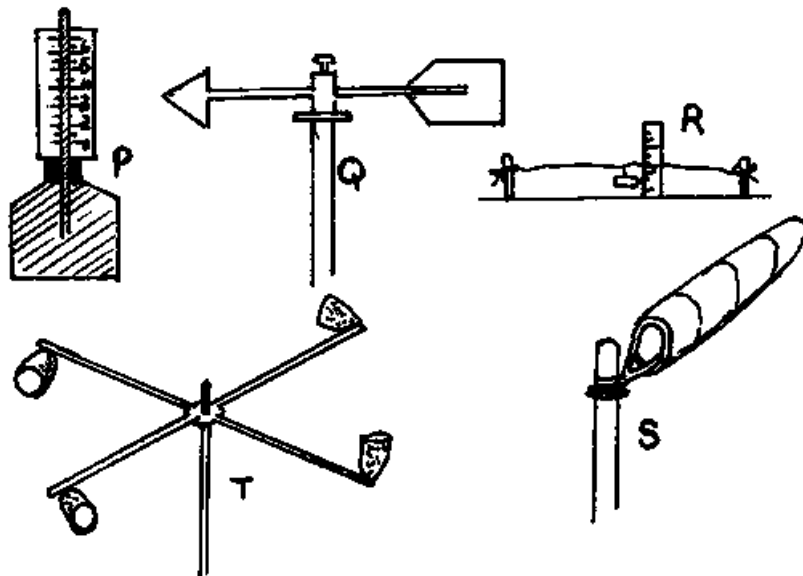
4. Look at the chart which gives the weather recorded by a Standard 8 class for a period of 5 days:

Day	Temperature	Humidity	Cloud/Sun	Rainfall
1	29°C	2 units	Half sun/half cloud	0 mm
2	19°C	8 units	no sun/full cloud	30 mm
3	20°C	6 units	half sun/half cloud	12 mm
4	28°C	2 units	sun	0 mm
5	30°C	1 unit	sun	0 mm

Which of the following conclusions are they most likely to have made?

- A. Low temperatures, low humidity and some sun bring rain.
- B. Low temperatures, high humidity and thick clouds bring rain.
- C. Low temperatures and low humidity *bring* clouds.
- D. High temperatures and high humidity bring some rain.

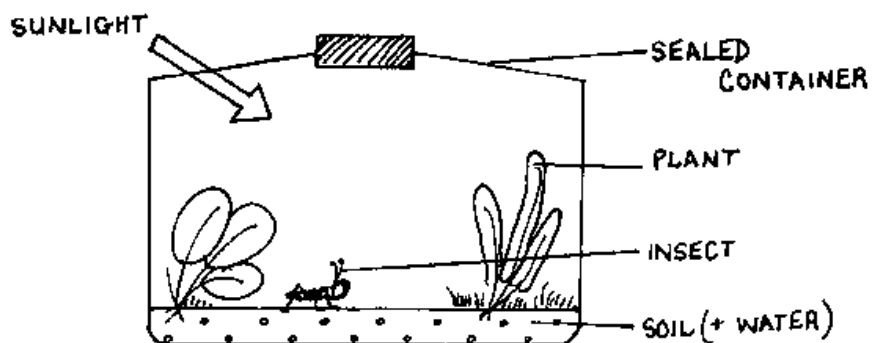
5. Look at these drawings of apparatus used to measure the weather:



Which two items of the above apparatus measure the SAME weather characteristic?

- A. P and R
- B. Q and R
- C. T and P
- D. Q and S

6. Study this arrangement of a sealed, large transparent container:



Select the MAIN REASON why these plants and the insect can live together (for some time) in the sealed container:

- A. The insect may eat the plants.
- B. The plants use the carbon-dioxide.
- C. The plants recycle oxygen and water.
- D. The plants use water.

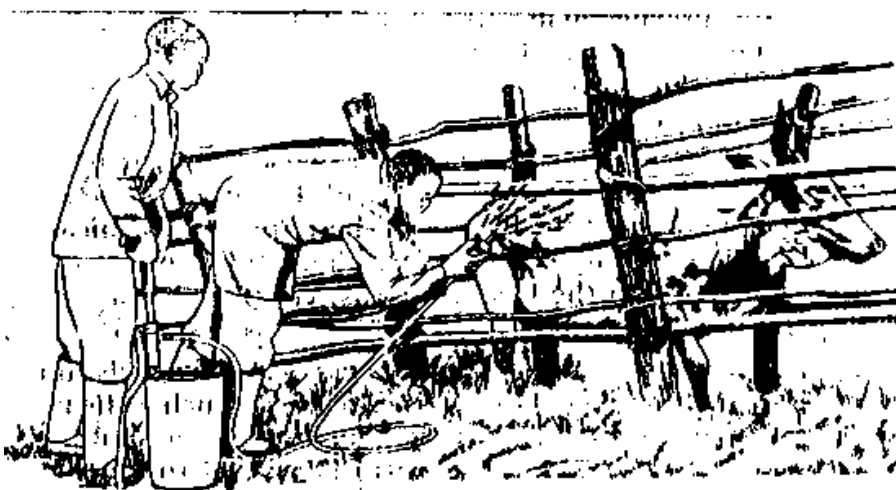
7. The list below gives the methods commonly used to control soil erosion

- i) mulching
- ii) planting vegetation
- iii) constructing gabions
- iv) terracing

Which of these methods would BEST be used in an area with a gentle slope?

- A. (i) and (iv)
- B. (i) and (ii)
- C. (ii) and (iii)
- D. (ii) and (iv)

8. The diagram below shows Peter and Mary carrying out an activity to control animal parasites.



The parasite being controlled is most likely:

- A. Flea
- B. Tick
- C. Tsetse fly
- D. Wireworm

9. The leaves of most tea plants on a farm look pale yellow between the veins while the veins themselves are green. What is wrong with the tea plants?

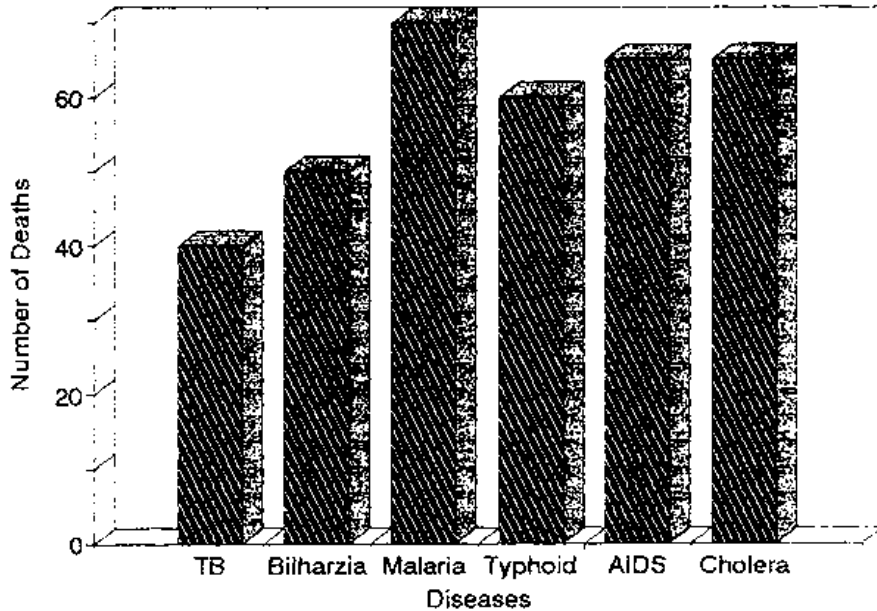
They are:

- A. deficient in nitrogen.
- B. deficient in magnesium.
- C. attacked by tea thrips.
- D. deficient in calcium.

10. Which of the following sets of diseases is caused by the same type of organism?

- A. cassava mosaic, tobacco mosaic and blight
- B. blight, coffee berry disease and rusts
- C. black rot, maize streak and bacterial wilt
- D. blight, black rot and rusts

11. Study the bar graph below which shows the number of people who died from various diseases in country X and answer the question that follows:

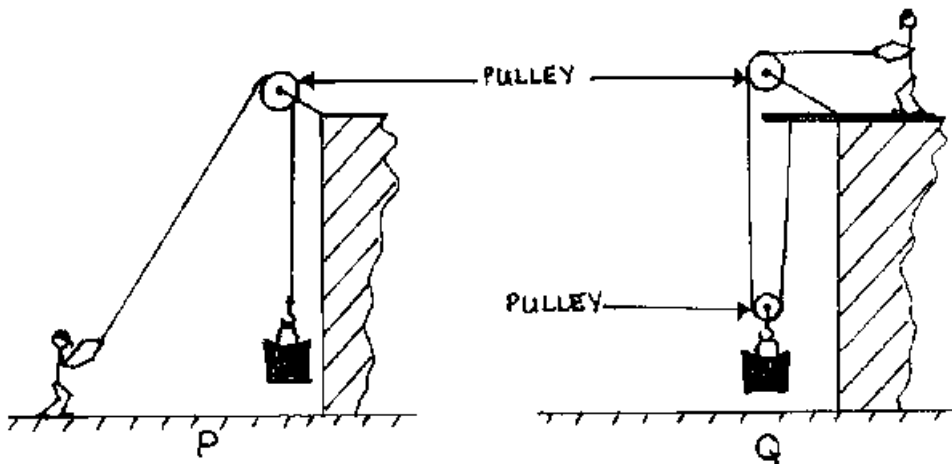


Deaths by Type of Disease

Which of the following is true according to the graph?

- A. Tuberculosis (TB) and Cholera killed 120 people.
- B. Bilharzia and Typhoid killed 120 people
- C. Malaria and AIDS killed 135 people.
- D. Cholera and malaria killed 140 people.

12. Diagrams P and Q show two men lifting two pails of equal weight up to the ceiling.



In which picture does the man have to pull less hard?:

- A. P because he is on the ground.
- B. Q because he is on the ceiling.
- C. Q because he applies less effort.
- D. P because he is pulling downward.

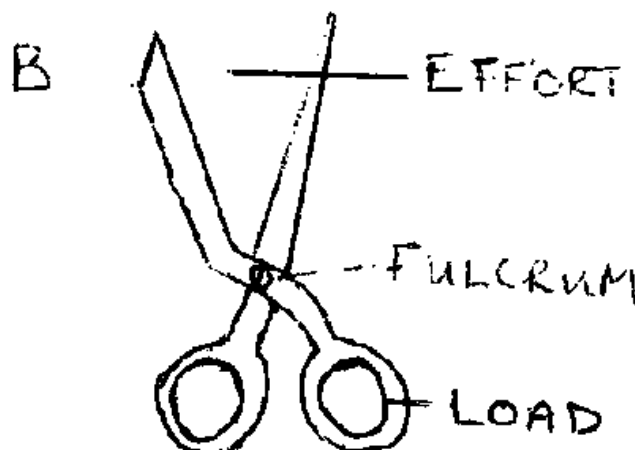
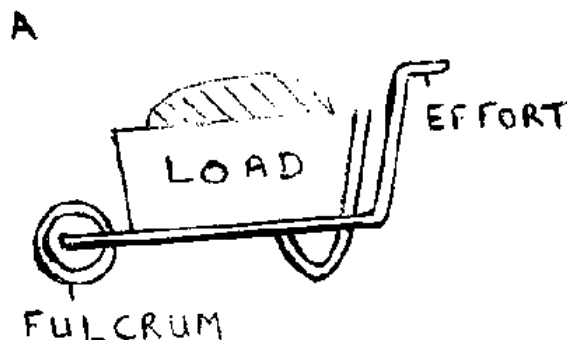
13. The table below shows an egg production record on Mr Kamau's farm in a certain week.

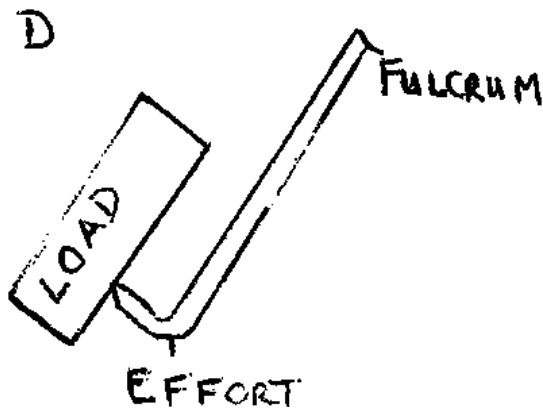
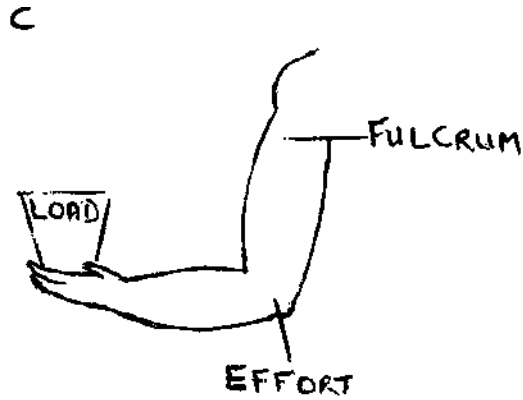
Day of the Week	Mon	Tue	Wed	Thur	Fri	Sat	Sun
Number of Eggs	60	56	65	63	67	55	60
Number of Damaged Eggs	7	2	11	9	12	0	6

If all the eggs were sold, on which days of the week did Mr Kamau get the highest returns?

- A. Monday and Friday
 - B. Monday and Tuesday
 - C. Thursday and Saturday
 - D. Friday and Saturday
14. What would be the main aim of a farmer who plants maize in the first season, beans in the second season, and potatoes in the third season in his shamba?
- A. To improve crop yield from an unproductive shamba.
 - B. To provide the farmer with a variety of crops.
 - C. To maintain a good crop yield without loss of soil fertility.
 - D. To select a crop most suited to the season.

15. In which of the following lever systems are the load, effort and fulcrum labelled correctly?



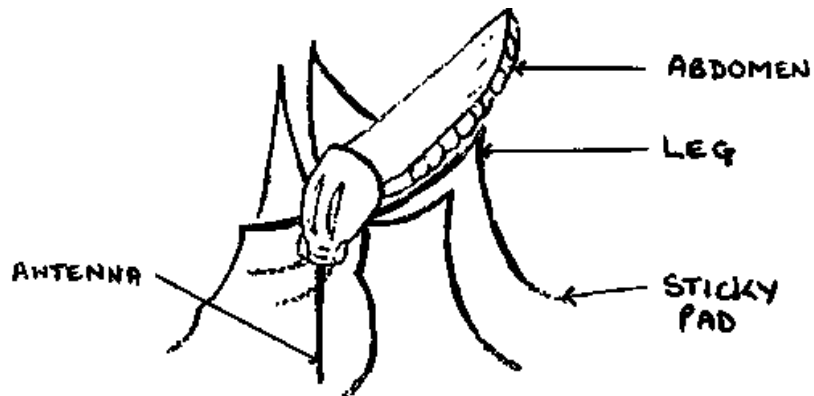


- A.
- B.
- C.
- D.

16. Maryam visits a farm where livestock are reared. She finds some of the cattle have watery eyes, hair loss on the tail end, and the abdomen and chest are swollen. Which disease are these animals likely to be suffering from?

- A. Anthrax
- B. East coast fever
- C. Nagana
- D. Foot and mouth

17. Below is a diagram of a certain insect. Which part is wrongly labelled?



- A. antenna
- B. leg
- C. abdomen
- D. sticky pads

18. Below is a record of sales and expenses for Tofila's farm in November 1993

SALES		
Date	Details	Amount
Nov 1	Sale of rabbits	600.00
Nov 12	Sale of carrots	350.00
Nov 25	Sale of Tomatoes	355.00
EXPENSES		
Date	Details	Amount
Nov 5	Wages	400.00
Nov 12	Pesticides	140.00
Nov 25	Fertilisers	800.00

Which one of the following conclusions can correctly be made from the record?

- A Tofila made a loss in November
- B. Tofila made a profit in November
- C. The sale of tomatoes could pay for fertilisers
- D. Tofila neither made a loss nor a profit

19. Which of the following farm information is correctly matched?

	Farm Information	Farm Record
A	Types of diseases and parasites contracted by farm animals	Marketing
B	Prices of farm products	Health
C	Sales and expenses	Inventory
D	Farm yields	Production

- A.
- B.
- C.
- D.

20. Below are a farmer's income and expenditure records for the month of January, 1993:

Date	Details	Shs
18-1-93	Paid for clearing field	830.00
20-1-93	Paid for ploughing	1050.00
25-1-93	Bought 50 Kg maize seed	2000.00
26-1-93	Sold chickens	2300.00
28-1-93	Sold eggs	1000.00
29-1-93	Sold tomatoes	580.00

What conclusion can be made from these records?

- A. The farmer made least profit from the sale of chickens during the month of January 1993
- B. The farmer made more profit from the sale of eggs than from the sale of tomatoes during January 1993
- C. The farmer neither made profit nor loss during the month of January 1993.

D. The farmer incurred loss during the month of January 1993

21. Which of the following things are interdependent?

A	B	C	D
Cattle	Pitcher Plant	Locust	Oxen
Tickbirds	Weaver Bird	Ants	Farmers

- A.
- B.
- C.
- D.

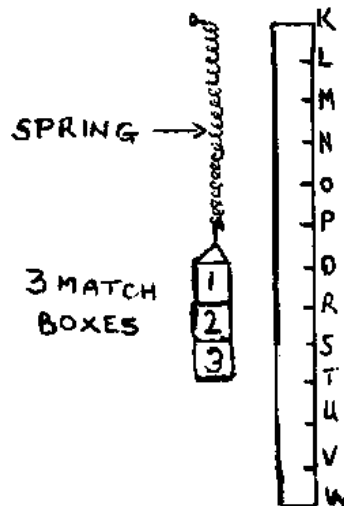
22. Which of the following shows the correct order of changes in state when water at 20° C was first cooled to -10° C and then heated to 100° C?

- A. Liquid → gas → solid → liquid
- B. Liquid → solid → gas → solid
- C. Liquid → liquid → liquid → liquid
- D. Liquid → solid → liquid → gas

23. A measuring cylinder was filled with water to the 30 cm³ mark. A glass cube of sides of length 2 cm was then lowered into the measuring cylinder. What was the final reading on the measuring cylinder:

- A. 8 cm³
- B. 28 cm³
- C. 32 cm³
- D. 38 cm³

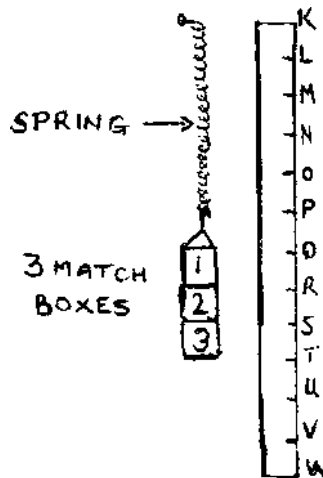
24. A stone was slowly lowered into water in a measuring cylinder marked in cm³. The water level was originally at the 15 cm³ level. The diagram shows the stone in its final position.



The volume of the stone was

- A. 10.0 cm³
- B. 7.5 cm³
- C. 22.5 cm³
- D. 15.0 cm³

25. A spring of length KM will stretch to the position P when three identical matchboxes are hung at one of its ends as shown in the diagram



Which of the following statements points out what is likely to happen when a total of five identical matchboxes are hung at the end of the spring? It will stretch to position:

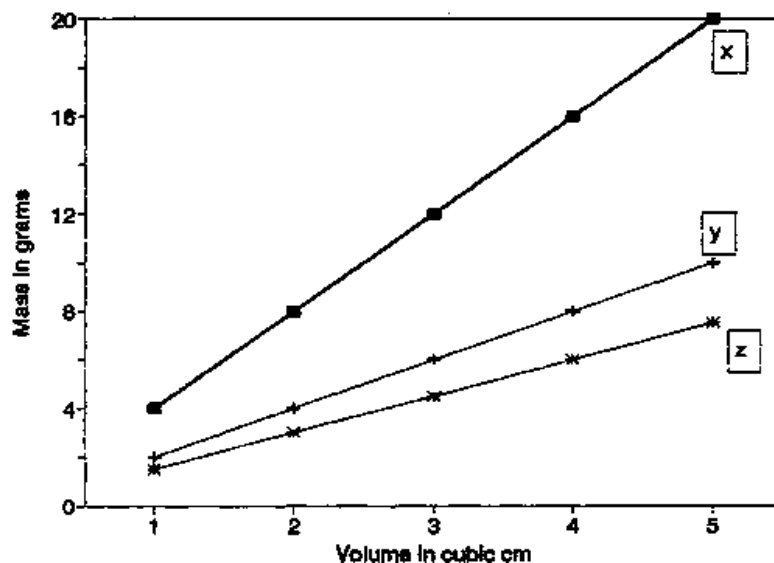
- A. P
- B. T
- C. U
- D. V

26. Peter noticed that in his maize shamba some plants had large holes on the leaves and stem. The maize plants had most likely been attacked by:

- A. Maize weevil
- B. Grasshopper
- C. Aphid
- D. Stalkborer

27. Below are graphs showing the relationship between mass and volume of three different liquids X, Y Z. The formula for density is

$$\text{DENSITY} = \frac{\text{MASS}}{\text{VOLUME}}$$



Mass of Liquid by Volume

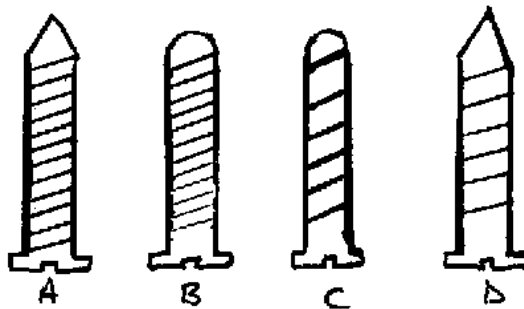
These three liquids were poured into one bottle shaken and allowed to settle as shown below.



Choose the correct statement below

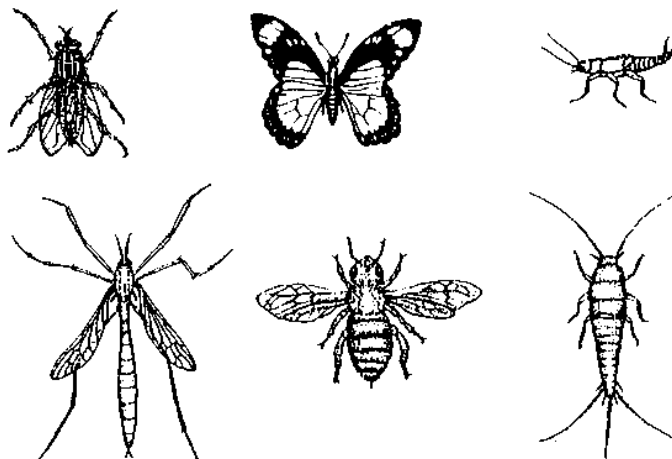
- A. Liquid Z was found in the middle
- B. Liquid X was found at the top
- C. Liquid Y was found at the bottom
- D. Liquid X was found at the bottom

28. With the help of a screw driver you are required to drive four wood screws given below into a piece of wood. Indicate below which screw would require least effort to go into the wood.



- A.
- B.
- C.
- D.

29. Observe the animals represented by the diagrams below and choose the statement that applies to all of them.

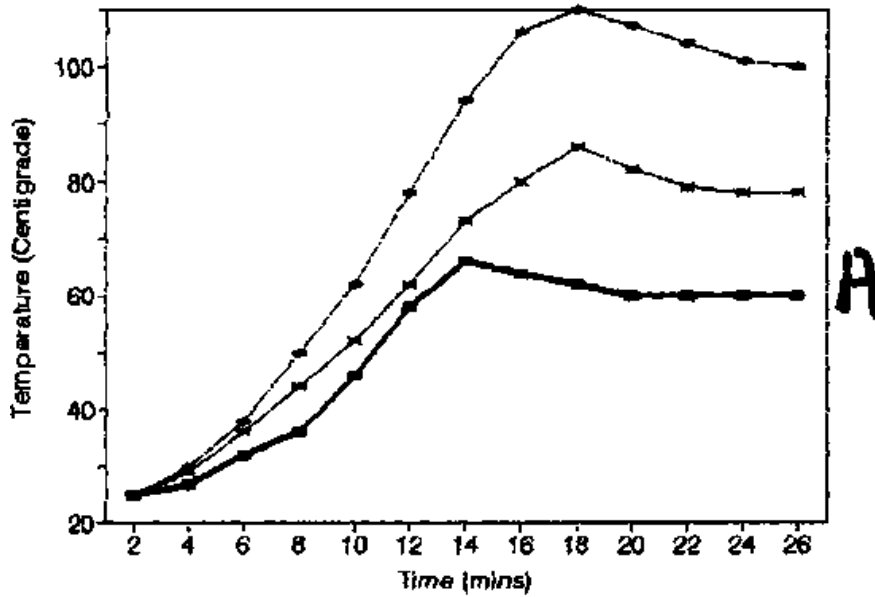


- A. The wings can be easily seen.
- B. The legs are longer than the body.
- C. There is a pair of antennae on the head.
- D. The abdomen has hair-like structures at the end.

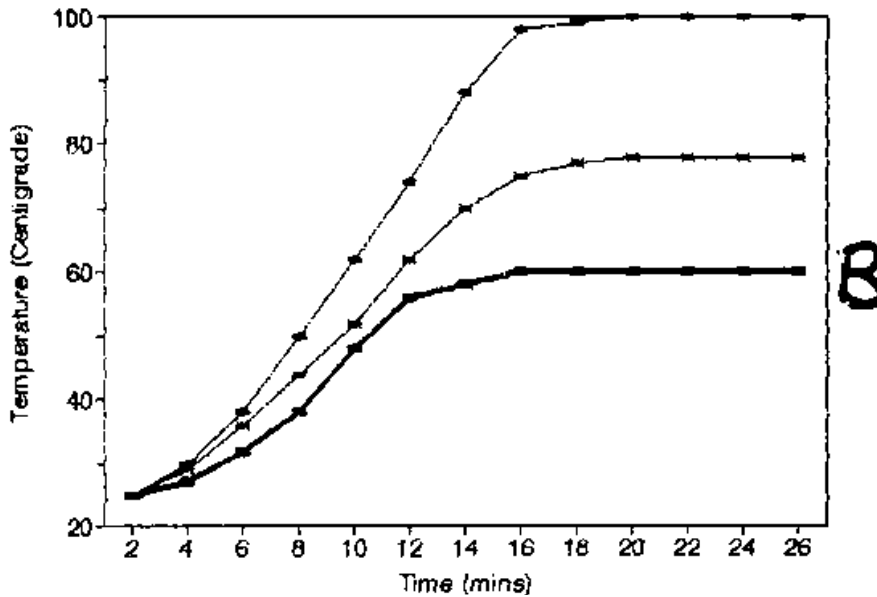
30. Pupils from Bemba Primary School heated three liquids separately and took the temperature of each every two minutes. After boiling they continue to take the temperature for about 10 minutes. The first liquid boiled at 78°C, the second at 100°C, and the third at 60°C

Which of the following graphs best show how the temperature of the liquids changed.

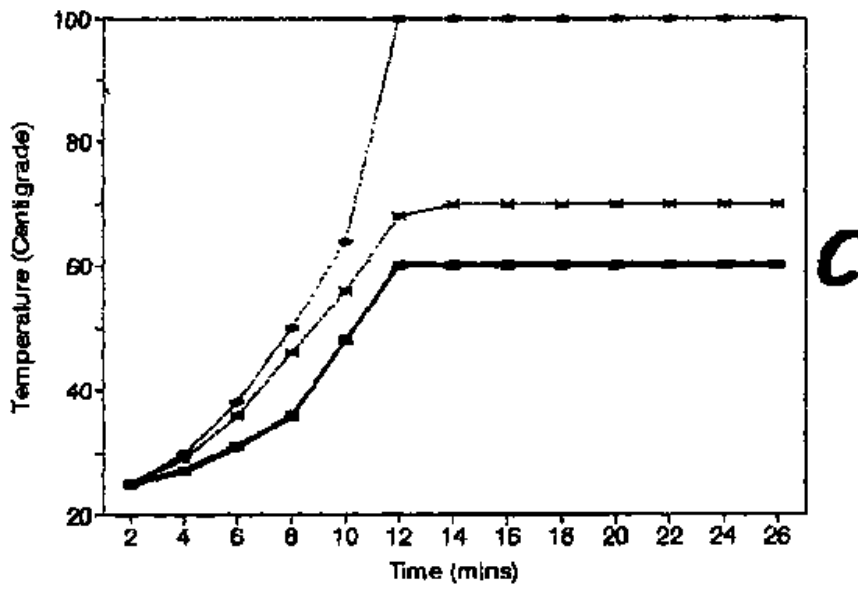
Rise in Temperature with Time



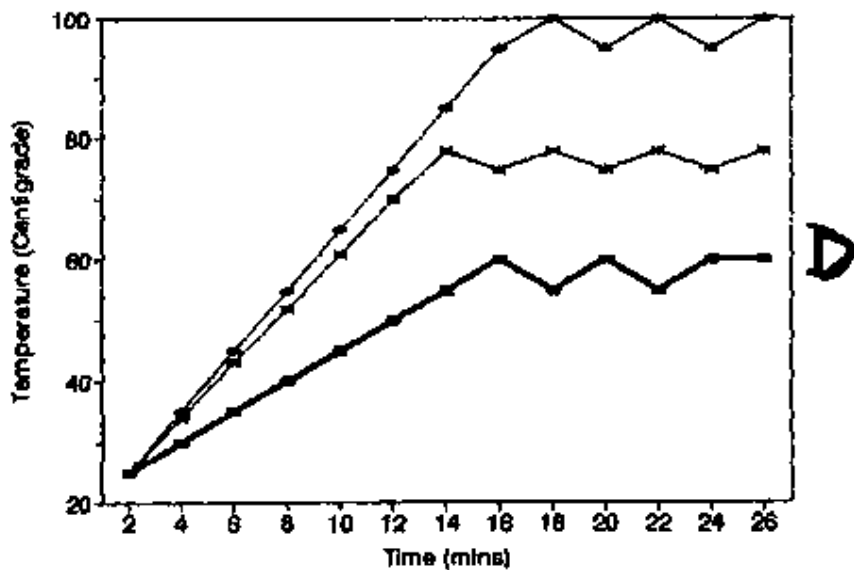
Rise in Temperature with Time



Rise in Temperature with Time



Rise in Temperature with Time



- A.
- B.
- C.
- D.

MARKING SCHEME FOR TRIAL TEST PAPER »SCIENCE & AGRICULTURE «

(1)	C	(16)	C
(2)	C	(17)	A
(3)	c	(18)	A
(4)	B	(19)	D
(5)	D	(20)	C
(6)	C	(21)	A
(7)	B	(22)	D
(8)	B	(23)	D
(9)	B	(24)	B

(10)	B	(25)	D
(11)	C	(26)	D
(12)	C	(27)	D
(13)	D	(28)	A
(14)	C	(29)	C
(15)	A	(30)	B

SUMMARY	
Key	Frequency of Items
A	5
B	7
C	10
D	8
Total	30

Assignment 5: Item Prediction

Look at the reviewed and edited 30 items of the final test paper. Before you start the trials in school try to assess the items in advance, and compare your predictions later on with the test results after the analysis of the data.

- Which items do you expect to be easy for the pupils?
- Which items do you expect to be difficult for the pupils?
- Which items would favour girls?
- Which items would favour boys?

- List those items which would discriminate well.
- List those items with the strongest distractors.

Assignment 6: Administration of Test

- Groups 1, 2, 3 split into two sub-groups.
- Each sub-group nominates test administrator.
- Test administrators meet to develop **brief note**:
 - How should the test be introduced?
 - What instructions should be given?
 - Time limit?
 - Queries?
 - Collection of papers.
- Other group members meet in group 1, 2 or 3. Design **interview agenda** for school visit.
 - Interview questions for Science teachers based on last year's examination paper, school based testing practices and assessment. Include some discussion of performance differences and difficult types of item.
 - Notes to review students' books.
 - Notes to observe facilities, especially textbooks, and school based tests.
 - Assume you have about 45 minutes to interview and 45 minutes to observe.

Info 6: Summary of Item Analysis Procedures

1. The analysis of multiple-choice items is usually undertaken for one of several reasons:
 - A pre-test analysis is undertaken on a representative sample of candidates to determine which items have appropriate difficulty and discrimination.
 - A post-test analysis is used to establish whether items have performed acceptably in terms of difficulty and discrimination.
 - Analyses are needed of particular groups of candidates (e.g. girls, rural and urban) to decide what differences there are in performance and which kind of items produce most of the variation in performance.
2. Item difficulty (D) is established by calculating the proportion of candidates obtaining a correct response to each item or to the test as a whole. The simplest formula for this is

$$D = N_c / N_t$$

where N_c = Number getting the correct response, and
 N_t = Number of candidates responding to the item.

3. For norm-referenced tests, where the performance of each candidate is compared to that of others (and the expected distribution usually approximates a normal curve), it is desirable that average difficulty is about 0.5. This allows the maximum variation between candidates to occur. If most items are very easy $D > 0.70$ or very difficult $D < 0.30$ it will be difficult to separate out the performance of candidates and the level of discrimination will be low.
4. It may be desirable to include more items with $D > 0.50$ than with less so that even poor candidates have the satisfaction of a numerical score that appears to approach 40 to 50%. The decision on the average D for a test is ultimately arbitrary, but it should not be so high that it seriously reduces discrimination levels in a norm-referenced test.
5. Items should be selected for their content validity. The value of D should be a secondary consideration. The value of D is not directly related to the cognitive level of items - e.g. recall items can have high or low D as can application items.
6. Criterion-referenced tests (where performance is compared to a criterion performance represented by the test item) are not concerned with discrimination across the range of scores. It is therefore inappropriate to select items for these tests on the basis of the "D"-value. The logic is that it is the criterion that is defined first and candidates are invited to achieve successful performance. If most do not, the criterion may be revised and vice versa. Mastery learning criterion-referenced tests assume that most will be able to master the performance and the value of D will be high.
7. Discriminating power refers to the ability of an item to separate candidates in a distribution that correlates well with performance on a complete test (or sub-test). High scoring candidates overall should score well on each individual item. If they do not, good candidates are performing poorly on an item and the reasons need investigation (e.g. they may be confused by ambiguous wording that less able candidates fail to perceive; the item may measure a different trait than the overall test).
8. The discrimination power (DI) of an item can be calculated simply by dividing candidates into quartiles (top 25%, upper middle 25%, lower middle 25%, lowest 25%) and applying the formula below:

$$DI = (N_u - N_l) / N_t$$

where N_u = Number getting the correct response in the top 25%,
 N_l = Number getting the correct response in the lower 25%,
 N_t = Total number of correct responses in upper and lower groups.

Alternative methods divide candidates into two or three groups rather than four.

9. When DI is positive and > 0 the item is discriminating in the same way the test does. When it is negative and $D < 0$ the item is penalising those who score highly overall. When D is close to zero no useful discrimination is occurring.

10. A DI > 0.3 is usually considered acceptable for a well behaved item. If DI is lower the item should probably be revised. Where DI is negative it may be measuring a skill or cognitive process unrelated to other items within the test.

11. Items can have an acceptable DI but have little or no content validity. The latter has to be assured.

12. Criterion-referenced tests need to discriminate in a different way to norm-referenced tests. With norm-referencing good discrimination is desirable across the whole range of marks. With criterion-referencing it is only significant at the level of the criterion performance - much better or much worse performance is not of concern for grading.

13. Distractors are those alternatives which are incorrect for each item. Analysis of the D and DI values of distractors can indicate whether

- many candidates chose one distractor (in which case it may be misleadingly worded);
- particular types of candidate are attracted to a distractor (in which case one group of candidates may be disadvantaged);
- no candidates chose a distractor (in which case it should be discarded);

14. Good distractors attract comparable numbers of lower scoring candidates for each alternative. Higher scoring candidates should be the minority of those who chose distractors.

15. A simple measure of the internal consistency of a test is to calculate performance on different groups of items (split halves) and compare these. The most convenient approximation of this is to use the Kuder-Richardson formula known as KR21.

$$\text{Reliability (KR21)} = \frac{k(1 - (m(k - m)/ks^2))}{k - 1}$$

k = number of test items

m = mean score (arithmetic mean)

s = standard deviation of test scores

16. KR 21 tests if all items measure the same thing (homogeneity). It is roughly equivalent to split-half tests of reliability where performance on half the items is compared to performance in the other half. It is not appropriate for tests where not all candidates complete all items (speed tests). The higher KR21 is the more consistent the test items.

17. Item analysis has to be undertaken for a purpose (item selection, post-test assessment, performance analysis). This should be decided at the outset.

18. The content validity of items is critical and consideration of this should precede consideration of D and DI.

19. Item analysis should be approached systematically and records kept of the results of each analysis so that it may be referred to subsequently. Item content should be recorded with item statistics.

Assignment 7: Preliminary Analysis of Data

- Give each candidate a number (e.g. School 1/1-50, School 2/1-63, School 3/1-48 etc.).
- For each school split papers into boys and girls.
- Mark whole set recording total score for each candidate.
- Note average score of boys and girls.
- Arrange papers in order of total score and separate into two halves (to get high and low scoring groups).
- Mark each item by counting number of A/B/C/D-responses in each group.
- Work out overall facility value

$$F = \frac{N_{\text{Correct}} \times 100}{N}$$

- Work out discrimination index for key and for distractors

$$D = \frac{N_U - N_L}{N/2}$$

(number of correct answers of top candidates (N_U) minus number of correct answers of bottom candidates divided by half of the total number of candidates ($N/2$))

- Work out **F** and **D** for boys and girls separately.

SCHOOL NAME

Item	Options	Facility Value	Discrimination Index	Facility Value	
				(Female)	(Male)
	A B C D	F	D	FF	MF
01	N_U				
	N_L				
02	N_U				
	N_L				
03	N_U				
	N_L				
04	N_U				
	N_L				
05	N_U				
	N_L				
06	N_U				
	N_L				
07	N_U				
	N_L				
08	N_U				
	N_L				
09	N_U				
	N_L				
10	N_U				
	N_L				
11	N_U				
	N_L				
12	N_U				
	N_L				
etc.					

N_u = number of upper group achievers
 N_l = number of lower group achievers

Info 7: A Simple Measure of Reliability

$$\text{Reliability (KR21)} = \frac{K(1 - (M(K - M)/KS^2))}{K - 1}$$

K = number of test items (30)
M = mean score (arithmetic mean)
S = standard deviation of test scores

- KR 21 tests if all items measure the same thing (homogeneity).
- Roughly equivalent to split-half tests of reliability.
- Not appropriate for speeded tests (if a larger number of students do not complete the test).
- Not an indication of constancy of score over time.

Note: See also Info 6, Nos. 15 & 16.

WORKING GROUP RESULTS

Outline for Interview with Science/Agriculture Teachers

WORKSHOP on»WRITING OF TEST ITEMS FOR PRIMARY SCHOOL LEAVING EXAMINATIONS«- Nyeri/Kenya, 1994

GROUP II

Name of School:

1. General Feelings about the 1993 Examination Paper

1. Level of difficulty
2. Language used
3. Accuracy of diagrams
4. Syllabus coverage
5. Time allocated to exams
6. Clarity of instructions
7. Gender or area bias
8. Chances of questions used in teaching being used in the National Examination
9. Who sets this exam, do you get a chance to be involved?

2. School-Based Assessment

1. How often do you test?
2. Who sets these tests?
3. Are there any exchanges between teachers and schools?
4. Are you given any training?
5. At what level of difficulty and on what sections do you test?

6. Do you evaluate your tests?
7. Do you have resources to help you set tests?
8. What is the purpose of this testing?
9. Do you keep records?
10. Are the results fed back to the pupils or parents (do you have prize givings)?
11. Do you institute remedial work to help those who are not passing?
12. What types of questions do you set, i.e. multiple-choice or longer answers (try to get a sample)?
13. What problems do you encounter with different methods?

3. Books

Ask these questions first about text books and then about the pupils' exercise books.

1. What type of books are used?
2. How many books?
3. Do you have a resource centre?
4. Do the pupils own their books or do they belong to the school?
5. Do the pupils value their books?

Now with regard to pupils' exercise books

1. How are the notes laid out?
2. How do you control the quality of the diagrams?
3. How frequently do you inspect the books?
4. How frequently are you inspected?
5. What problems do you encounter e.g. pupils losing books, copying?
6. Any other comments?

7. Select pupils' exercise books at random and study them carefully,

paying special attention to the following:

- Are the exercise books kept regularly?
- Do pupils describe things in their own words?
- Are informations kept uniformly in all books?
- Do the books mainly contain copy work from chalkboard?
- Check topics covered with syllabus for Std. 7.

WORKSHOP on»WRITING OF TEST ITEMS FOR PRIMARY SCHOOL LEAVING EXAMINATIONS«- Nyeri/Kenya, 1994

GROUP III

Outline for Interview with Science/Agriculture Teachers

Name of School:

Sex of interviewee:
Teaching experience:
Grade taught:
Teacher's qualification:

Number of Science & Agriculture teachers:
Number of Science & Agriculture periods per week:

• INTERVIEW QUESTIONS BASED ON THE 1993 KCP SCIENCE AND AGRICULTURE EXAMINATION

1. Was the syllabus adequately covered by the examination?

Yes No

2. Did the examination meet the curriculum objectives?

Yes No

3. To what extent was the Science syllabus covered?
(Last year before they sat for the final exam?)

Less than half Three quarters
Half More than three quarters

4. Would you say last year's paper was easy or difficult?
Give reasons.

5. Which questions do you think were difficult?
Give reasons.

6. Which questions do you think were easy? Give reasons.

7. Are you happy with the multiple-choice type of questions in the exam?

Yes No

Give reasons.

• QUESTIONS ON SCHOOL-BASED TESTING PRACTICES

1. How often do you test your pupils?

2. How do you generate your test items?

KCPE past papers Set own questions
Other methods (specify)

3. Do you meet any problems when administering school-based tests?

Yes No

4. What do you think is the root cause of these problems?

5. How have you tackled these problems?

6. To what use do you put the results of the tests?

• QUESTIONS RELATED TO PERFORMANCE DIFFERENCES

1. How was the performance of your pupils in the 1993 KCPE Science paper?

Overall Good Average Bad

2. How do you compare the performance of

- boys and girls?
- rural and urban pupils?

3. What suggestion would you make so as to attain better performance of your pupils in the KCPE Science and Agriculture papers?

• **BOOKS AND EQUIPMENT**

1. Does the school provide textbooks to pupils?

Yes No

If yes, are they adequate?

2. Are the books suitable to the syllabus?

Yes No

3. What equipment does the school have for Science and Agriculture teaching?

• **NOTES TO REVIEW PUPILS' NOTE BOOKS**

The team should review 5 pupils' Note Books selected randomly from Std. 8 and note

1. Topics covered
2. Adequacy of notes
3. Relevance
4. Accuracy of notes, diagrams etc.
5. Evidence of teacher review of notes

Reports on Interviews in Primary School

WORKSHOP on»WRITING OF TEST ITEMS FOR PRIMARY SCHOOL LEAVING EXAMINATIONS«- Nyeri/Kenya, 1994

GROUP IA

1. Introduction

This is a report of an interview held at **Kangocho Primary School** on 22nd March, 1994. The Science/Agriculture teacher was not available since he had gone to hospital. Some teachers, a Std. 6 Science teacher and the headmaster of the school were interviewed instead. The school has six subject panels for Languages, Mathematics, Geography/History/Civics/Religion, Home Economics/Business Education, Art & Craft, Science/Agriculture, each comprising six teachers on average.

2. KCPE 1993 Science/Agriculture

The paper was considered to have fair and adequately covered the syllabus. Boys performed better than girls. Teachers are fairly happy with the KCPE format but some feel that multiple-choice items deny pupils chances to reason and express themselves. KCPE results of a given year are analysed to ensure that weak areas are given adequate attention in the teaching of candidates for the following year.

With regard to pupils' preparation for the exam, the pupils are adequately prepared through zonal, divisional and district level examinations. In Standards 4 to 8 pupils are given nine examinations per year and during the final year (Std. 8) the third term is devoted to KCPE

past papers. Pupils write the exams and difficult topic areas are identified and discussed by the teacher(s) and pupils.

3. Teaching/Learning of Science/Agriculture

Teaching is rather theoretical due to lack of equipment. The school does not have a shamba for practical agriculture and there is no science room. The science equipment is borrowed from the nearby secondary school, but the school has spring balances, thermometers, torch cells and circuit wires (leads).

The set books are those recommended by the Kenya Institute of Education. Two pupils share a book. Textbooks are given to the pupils in January and collected in December of each year. Some parents buy textbooks for their children as encouraged by the school management. Textbook needs are discussed during staff meetings and the money available, raised through Harambee meetings and donations, is used according to decisions of the staff. Examination of note books showed that pupils are given some notes and a few exercises to do.

4. Test Construction/Generation

Teachers have an idea of test construction as taught during methodology courses at Teachers' Training Colleges; but there is no course in Measurement and Evaluation per se at Teachers' Training Colleges.

The setting of zonal, divisional and district level examinations is done by subject panels which use the KCPE format but panelists are not experts in setting questions.

5. General Comments

The education system in Kenya seems to be exam driven since it is highly competitive. Besides grading and ranking pupils' performance, schools and districts are ranked. This in effect forces teachers to teach to the exam. Syllabuses are rushed through to ensure coverage so that pupils are prepared for the exam. One would only feel sympathy for the pupils as regards the number of examination papers they have to go through. Somehow the pupils tend to cope with the situation.

WORKSHOP on»WRITING OF TEST ITEMS FOR PRIMARY SCHOOL LEAVING EXAMINATIONS«- Nyeri/Kenya, 1994

GROUP III B

1. Introduction

Three members of the group interviewed the headteacher and three Science and Agriculture teachers of **Nyamachaki Primary School** on 22nd March, 1994 between 11.30 a.m. and 12.45 p.m. The interview was carried out by following guiding questions which had been prepared the previous day.

The purpose of the interview was to collect information about the school concerning its background history, current number of Science and Agriculture teachers, time allocation and the teaching of Science and Agriculture subjects, availability of facilities and teaching/learning materials, 1993 Kenya Certificate of Primary Education (KCPE) Science and Agriculture examination and school-based testing procedures practised.

2. Findings of the Interview

2.1. Background History

Nyamachaki Primary School is located a few kilometres from Nyeri town in Central Province. The school was established in 1933. It started with 560 pupils, but the number has risen up to 1,169 pupils currently. Most of the pupils in the school are residents of Nyeri town and from surrounding areas. However, there are some pupils who come as far as 10 km from the

school. These pupils travel by bus to and from the school every day.

2.2. Current Number of Teachers

The school has a total number of 33 teachers out of whom six are men and the rest are women.

Out of the 33 teachers, eight teach Science and Agriculture subjects in Std. 3-8. Three of them teach in Std. 7 and 8. They are trained grade PI teachers with a teaching experience of more than four years.

Of the three Science teachers teaching in Std. 7 and 8, two are females, one is male.

2.3. Time Allocation and the Teaching of Science and Agriculture

The Science and Agriculture subjects are each allocated three periods per week. Interviewed teachers remarked that time allocated for Science was not enough, but the time for Agriculture was sufficient. They said they used some of the Agriculture periods for teaching Science.

Pupils of Std. 4-8 are allocated 2.15 hours per week in the school timetable for remedial instruction and preparatory work. If this time is well utilized by pupils and teachers it is likely to improve pupils' progress and performance. Teachers mostly use the lecture method of teaching. Very few practicals and demonstrations are done during teaching.

2.4. Availability of Facilities and Teaching/Learning Materials It was learnt from teachers during the interview that:

- the school has a workshop room for Art and Craft and Home Science practical instruction;
- the school has buildings that are enough for 24 streams/classes;
- out of the 24 classes, 9 of them are semi-permanent structures. The walls and roofs of these structures were constructed with corrugated iron sheets;
- pupils' textbooks for Science and Agriculture subjects are not enough. The ratio is about one book for ten pupils;
- the books are suitable for the syllabuses of Science and Agriculture subjects;
- there are insufficient numbers of teachers' guides and syllabuses. There are only three syllabuses for Science and Agriculture for the whole school;
- there are no Science kits for practical instruction;
- the school has a demonstration plot but currently it is not used for offering practical instruction for Agriculture lessons.

2.5. 1993 KCPE Science and Agriculture Examination

The following were findings from the interview with regard to last year's KCPE Science and Agriculture examination:

- The syllabus was adequately covered by the examination.
- The examination adequately met the curriculum objectives of the subjects.
- Before Std. 8 sat for the examination, more than 3/4 of the syllabuses for the subjects had already been covered.

- Interviewed teachers indicated that the following items¹ were:
 - very difficult: Items No. 22 and 42;
 - difficult: Items No. 3, 4, 10, 16, 18, 21, 29, 30, 32, 35, 37, 48, 57, and 60;
 - moderately difficult: Items No. 2, 5, 11, 13, 17, 19, 25, 28, 31, 33, 47, 49, 52, 53, 55, and 56;
 - easy: Items No. 1, 6, 7, 8, 9, 12,14,15, 20, 23, 24, 26, 27, 34, 36, 38, 39, 40, 41, 43, 44, 45, 46, 50, 51, 54, 58, and 59.

¹ See: Vol. 2 - Assessment of Science Mid Agriculture in Primary Schools in Africa; 12 Country Cases Reviewed

- Last year's exam was generally of the right standard.
- Interviewed teachers explained that some of the items were:
 - difficult because they required memorizing complex subject matter; interpretation of complex material asked in the items and involved calculations; also pupils had not performed many experiments; some of the examined material on diseases had not been taught because the syllabus does not specify which diseases have to be taught;
 - easy because they were short and of recall type; were more or less related to pupils' environment and experiences or had diagrams which were easy to analyse.
- Boys performed better than girls and most of those who performed poor were girls.
- Urban pupils performed better than pupils who came from rural areas.
- The following suggestions were put forward so as to attain better performance of pupils in the KCPE Science and Agriculture exams:
 - provision of adequate and accurate materials for undertaking practical activities (they are not provided so far);
 - provision of enough teaching/learning materials;
 - to have 5 periods per week for Science teaching;
 - teachers to be given seminars on topics in which they have low competence in teaching them effectively.
- The average score of the school last year was 431.67 out of a total score of 700. Hence the school was among the best schools in the District. The average score for the school has never gone below 404 for the past three years.

2.6. School-Based Testing Practices Undertaken

It was found out from the interview that:

- the teachers give quizzes after covering every topic. The quizzes are prepared by the teachers themselves using the syllabus and other course materials;
- pupils of Std. 6-8 are given 3 external exams per term;
- results of test/exam scores are recorded and used for guiding the teacher in undertaking remedial teaching;

- no problems are encountered in administering school-based tests apart from inadequate availability of stationery;
- pupils are prepared for exams by being drilled in answering past KCPE papers and other examination papers.

Reports on Test Administration

WORKSHOP on»WRITING OF TEST ITEMS FOR PRIMARY SCHOOL LEAVING EXAMINATIONS«- Nyeri/Kenya, 1994

GROUP IA

School: **Kangocho Primary School**

Class Tested: Std. 8 (14 year olds)

No. of Pupils: 64 (32 boys + 32 girls)

Progress of test:

- (1) After 15 minutes, most candidates on No. 11.
After 30 minutes, most candidates on No. 20.
After 55 minutes, most candidates finished.

Candidates checked their answers once they had completed all items.

- (2) All candidates brought pencils and erasers; they appeared to have been briefed beforehand and were co-operative and well disciplined.
- (3) One should bear in mind we are testing long-term memory - the children were not asked to revise their Std. 7 Science and Agriculture knowledge; certain factual recall (diseases, symptoms, etc.) may be poor as a result.
- (4) All of the Std. 8 pupils in Kangocho Primary School wrote the test.
- (5) All candidates completed the test in the time allocated (1 hour).

WORKSHOP on»WRITING OF TEST ITEMS FOR PRIMARY SCHOOL LEAVING EXAMINATIONS«- Nyeri/Kenya, 1994

GROUP III A

School: **Mathaithi Primary School**

Introduction

The test was administered on the 22nd March, 1994. The pupils had just finished writing a KCPE Mock Examination Paper. They were 78 candidates and could not be fitted in one room.

General Instruction

The test administrator explicitly explained why the test was administered. It was administered to find out how good or bad the test was. It was not part of the mock exam they were going through. It was most unfortunate that the whole exercise coincided with the time when they were writing their mock. The test administrator went over the general instructions. They were told they would not be given separate answer sheets. They would circle the letter that

corresponded with the correct response. An example was demonstrated on the chalkboard. Rubbers were put on each desk just in case the candidates wanted to make a correction in their responses. Caution was made on number 25. They were asked to alter the distractors on that item and write those written on the chalkboard. Half the pupils were asked to go to another room. Mostly those who remained were boys. The starting time of the test was 10.10 a.m. and the finishing time was 11.10 a.m. The information was displayed on the chalkboard.

General Observations

Some candidates, mostly boys, took a shorter time. Others even went over the paper trying to make readjustments to those items they did not get the answers straight away. No candidate needed any help and explanation of any item. Pupils' discipline was fantastic.

Physical Condition of the Classroom

The classroom set up was conducive to class activities. It was very spacious. Most space was taken by the big tables the children were writing on.

Reports on Test Results in Schools

WORKSHOP on»WRITING OF TEST ITEMS FOR PRIMARY SCHOOL LEAVING EXAMINATIONS«- Nyeri/Kenya, 1994

GROUP IA

School: **Kangocho Primary School**

Number of candidates: 64 (32 male & 32 female)
Number of test items: 30, each scoring one point

Performance Statistics:

- Overall average: 16.56
- Average for males: 17.03
- Average for females: 16.09
- Range of scores: 11 to 26
- Standard deviation: 3.07 (corrected)
- Test reliability: 0.22 (KR21)

Item Analysis

1. Facility

- (1) Items within acceptable range of difficulty of 30-70%: 4-6, 12, 13, 17, 20, 22, 24, and 27 (10 items)
- (2) Items above acceptable range of difficulty: 1-3, 8, 11, 14, 15, 18, 19, 26, and 29 (11 items)
- (3) Items below acceptable range of difficulty: 7, 9, 10, 16, 21, 23, 25, 28, and 30 (9 items)

Note: See items in the TRIAL TEST PAPER, pp. 82-98.

Possible explanations of items above acceptable range of difficulty:

- Item 1: The candidates appeared to be well prepared for reading a graph.
Item 2: This was a simple knowledge (recall) question.
Item 3: This was within the candidate's daily experiences in agriculture.
Item 8: No comment as $F = 100$ (and $D = 0$).
Item 11: Same as for Item 1 above.
Item 14: $F = 70.31 \approx 70$, acceptable.
Item 15: Simple observation in everyday situations.
Item 18: »Sales«and»Expenses«were clues.

- Item 19: Very simple classification.
- Item 26: Same as for Item 3 above.
- Item 29: Simple observation.

Possible explanations of items below acceptable range of difficulty:

- Item 7: Candidates may not know the meaning of »gentle« and »steep« slopes.
- Item 9: Candidates may not be familiar with this type of deficiency disease in tea; inadequate coverage in textbooks.
- Item 10: Advanced classification required.
- Item 16: Not within the pupils' everyday experience.
- Item 21: The term »interdependent« misunderstood. (As predicted!)
- Item 23: Poor concept of volume and, possibly lack of equipment.
- Item 25: Information lacking in stem and of comprehension level calling for prediction.
- Item 28: The pointedness of the screws became the criterion rather than the threads.
- Item 30: Lack of equipment and possibly not knowing that boiling takes place at constant temperature.

2. Discrimination

Items displaying poor discrimination are:

- Item 1: Very easy for both upper and lower groups, therefore, no discrimination.
- Item 6: The majority of the upper group opted for the distractor. The diagram emphasized use of CO₂. The word »recycle« was unfamiliar.
- Item 8: Familiar to both groups.
- Item 11: As in Item 1 above.
- Item 13: Both groups found the calculation simple.
- Item 14: Option A because of its closeness to the key C attracted the upper candidates.
- Item 16: Both upper and lower groups were unfamiliar with the disease, hence choice of B which is familiar to them.
- Item 19: Borderline.
- Item 22: The topic was moderately understood by both groups.

3. Distractors

- Item 1: Distractors did not work - the question should have been pitched at a higher level.
- Item 2: Distractors for A and B are negative; possibly »incorrectly« was misunderstood. Option D did not work.
- Item 3: Distractors for B and D are negative; possibly lower candidates are not familiar with »shading«.
- Item 4: Item had too many variables for the lower group to establish the overall relationship.
- Item 5: Distractors B and C hardly worked; the instruments were obviously different in function. The key D and option A worked equally.
- Item 6: Distractor B with highest D (0.22) strongly attracted candidates in both groups not identifying the key.
- Item 7: The option for »constructing gabions« only appears once and in distractor C; this might have been an attraction for the lower group. The meaning of »gentle« was misunderstood hence D = 0 for distractor D.
- Item 8: Distractors did not work.
- Item 9: Distractor A attracted the majority of the candidates as they are familiar with the element Nitrogen. The negative Ds for options A and C are insignificant (difference of 1).
- Item 10: The variations are mainly due to guess-work.
- Item 11: Distractors A and B hardly worked. Distractor D functioned despite negative D.
- Item 12: The distractors worked for both groups more to the disadvantage of the lower group.
- Item 13: Distractors B and C hardly worked.
- Item 14: Distractors B and D hardly worked.

- Item 15: The lower group was attracted by option C, possibly thought movement involved the whole arm.
- Item 16: Both groups were attracted by distractors A and B possibly because they are familiar.
- Item 17: Distractor B did not work because it was obviously correct. Diagram misleading.
- Item 18: Distractor C did not work; distractors B and C attracted more of the lower group because of poor addition skills.
- Item 19: Distractors A and B did not work effectively.
- Item 20: Distractors A and D did not work. Distractor B was the most effective.
- Item 21: Distractor D was the strongest because of the misunderstanding of »interdependent«.
- Item 22: Distractors A and B worked well; C was weak (only attracted lower group).
- Item 23: Distractor A hardly worked. Weaker candidates opted for B (poor volume concept).
- Item 24: Distractor A hardly worked; C and D attracted the lower group.
- Item 25: Distractor A hardly worked; distractor C attracted most candidates.

WORKSHOP on »WRITING OF TEST ITEMS FOR PRIMARY SCHOOL LEAVING EXAMINATIONS«- Nyeri/Kenya, 1994

GROUP II B

School: **Kwanderi Primary School**

$$X_{all} = 15.073$$

$$X_B = 15.833$$

$$X_G = 14.0$$

$$\text{Standard Deviation} = 3.054 \quad \text{Reliability} = 0.777$$

Facility index	30 - 70%	=	4, 5, 13, 14, 15, 17, 20, 22, 24, 26, 27, 28
	> 70	=	1, 2, 3, 8, 10, 18, 19, 29
	<30	=	6,7,9,10,12,16,21,23,25,30

Boys scored well on 15 items.

Girls scored well on 13 items.

Boys = Girls 2 items.

Item Analysis

- Questions 1, 2 and 3 were very easy with facility indices above 80%. Maybe this was due to the fact that the purpose was to ease the candidates into the examination.

- Question 3

Distractor A did not work well because the discrimination index was 0.

- Question 4

Discriminated very well. All the distractors were quite reasonable.

- Question 5

Most candidates chose A; maybe diagram R did not appear familiar to them. The key has a negative discrimination index which means that $N_i > N_u$; the »wrong« students, those that performed badly on the whole test, got the item correct.

- Question 6

Distractor D did not discriminate at all. Most candidates chose option B. Maybe the word »recycle« was not familiar to the candidates.

- Question 7

To a majority of candidates D was the most popular, because they did not seem to know the difference between a gentle slope and just a slope. The word »gentle« did not guide them in any way.

- Question 8

Distractors A and B did not discriminate at all. This might have been due to the inability to classify the parasites into internal and external parasites and then to decide on the method used to control the parasite. It might be that the pupils did not thoroughly analyse the diagram. Discrimination index for B was negative, so the distractor did not work well.

- Question 9

The most popular distractor was D; the key was negative which means that it did not work well $N_i > N_u$.

- Question 10

There was clear evidence of guessing since all distractors were chosen. This could be due to the inability to classify diseases according to the causative agent. This was a very difficult question with facility index of 0.293, which is low.

- Questions 11, 18, 19, and 29 were quite easy with a facility index above 0.8.

- Question 12

C and D have positive discrimination indices. So they were behaving the same way as the total test. C was the key but D appeared correct to the candidates. This clearly shows that there are hardly any practical experiments carried out in the classroom.

- Question 13

Some candidates chose A. This could be due to the fact that they concentrated on the number of eggs produced, but did not take into account the number of broken eggs.

- Question 14

Most candidates either chose option B or D because the idea of conservation of soil fertility did not click, all they considered was crop type and season. However, the facility index was 0.61 showing that the question was not very difficult. In fact it proved to be difficult for candidates to evaluate each distractor and decide the main reason.

- Question 15

The question was not accurately read so that for distractors B and C the discrimination index was 0.

- Question 16

The discrimination index for B was positive, so $N_i > N_u$. Therefore, the distractor did not work.

- Question 17

Maybe the »antenna« was not clearly labelled, so candidates may have been affected in a way. All the distractors worked, in fact it was a relatively easy question.

- Question 20

Distractors C and D have discrimination indices which are positive meaning that $N_i > N_u$; therefore, the distractors did not work.

- Question 21

The key A has a negative discrimination index, so the distractor was poor. To the candidates D appeared to be the correct answer. It seems that the stem needed to be improved, because it was not quite clear what type of interdependence was required.

- Question 22

It was not a very easy question. The discrimination index for B was 0, so it did not discriminate at all.

- Question 23

Distractors A and D have positive discrimination indices; $N_i > N_u$.

- Question 24

It was clearly evident that the candidates had problems with reading the scale such that options A, B and D did not discriminate well. A and B both had positive discrimination indices and D did not discriminate at all.

- Question 25

Absence of practicals actually being conducted in class hampered the students' ability to score. Most candidates chose C; $N_i > N_u$. So the distractor did not work well.

- Question 26

This was a very difficult question for the candidates and there was a lot of guessing. For all the distractors DI was positive.

- Question 27

The graph proved difficult to the candidates such that discrimination index for C was 0.

Question 28

Most candidates chose option D because they concentrated on the tips of the screws not the number of threads. In fact the only variable should have been the number of threads and the tips should have been the same.

- Question 30

It was the most difficult question. Candidates failed to interpret the graphs.

2.3.2. Reviewing the Item Writing and Testing Process

The objective of the exercise on item writing, testing and analysis was to train participants in the construction and analysis of test items for Primary Science and Agriculture. This was approached through a practical task which entailed writing assessment items, constructing a trial examination paper, applying it to groups of pupils and analysing the results. This involved several steps which included:

Deciding a Specification for the Range of Items to be Constructed

Multiple-choice items are widely used in assessment at the end of primary school in many African countries. The Certificate of Primary Education (CPE) in Kenya consists entirely of multiple-choice items and pupils are used to the format. Many of the underlying principles of item construction and analysis can be demonstrated through the development of multiple-choice questions. For these reasons, and the limitations of time it was decided to focus on the construction of an instrument with about 30 multiple-choice items. This was a sufficient number for all participants to contribute to the process.

Since the instrument had to be piloted on Kenyan pupils the content for items had to be drawn from the Kenyan Science and Agriculture curricula. It was decided to try out the instrument with pupils entering Grade 8 and thus material from Grade 7 was identified as that which was most appropriate to use.

To focus the exercise specific areas of the curriculum were identified and several items were written relating to each. As the intention was to concentrate on developing items above the level of recall five skill areas were identified in which items could be produced. The areas and the content topics chosen are indicated in **Assignment 1**, Table of Specification - Topics by Skills.

The skills identified were some of those judged most important to develop at primary level.

They were chosen after discussion of the meaning of each which can be summarised briefly as:

Observation	=	skills of looking and seeing
Classification	=	skills of categorising cases according to rules
Measurement	=	skills of deciding how or what to measure
Recording Data	=	skills of deciding how to record data
Interpreting Data	=	skills of interpreting data e.g. graphically

Development of Test Items

Test items were developed by participants working individually after discussion of the approaches that could be used and the pitfalls to avoid. Kenyan curriculum materials in Science and Agriculture were made available. Each individual produced at least two items for comment and improvement by other participants.

The development process alternated between writing periods and collaborative discussion of items produced in small groups. This was the first level peer auditing process. Participants were then grouped into three working groups and each was asked to produce between 10 to 15 items after internal review. This resulted in the production of more than 40 items which were then produced in the form of a trial test paper. This paper was then reviewed and edited to select the 30 items required for the final version.

As an additional exercise participants were asked to predict which items would be easy or difficult, which would discriminate well, which would be the strongest distractor, and which would favour boys or girls.

Trial of Test Items in Schools

Six local schools were identified where the test could be applied. These included some schools ranked very highly in national lists and other schools that were typical of the Nyeri District. Six teams of participants were organized, one for each school. In each team one member took responsibility for administering the test and the others arranged to interview school staff about assessment issues. Test administration was standardised as much as possible. The test was administered to a whole class group in all the schools except for one where the whole year group was tested. The test was timed for one hour.

Test administrators and the participants who interviewed teachers were asked to write up their observations briefly to feed into the discussion of how the test performed.

Marking of Responses

Altogether 338 papers were analysed from the six school samples. Marking was organized in school groups and participants developed an agreed procedure. First scripts were sorted in to boys and girls and then marked. This enabled overall averages to be calculated and differences in average score between girls and boys to be noted. Facility indices for items could also be calculated. The papers were then ordered from the best to the worst and performance of the top and bottom halves was calculated to arrive at a simple discrimination index. The power of different distractors was also assessed. A grid for results was developed to help make the task simple and systematic as shown below.

SCHOOL NAME:

Item	Options	Facility	Discrimination	Facility for	
				Females	Males
	A B C D				
1	N_u N_l				
2	N_u N_l				
3	N_u N_l				
4	N_u N_l				

Facility values are given by:

$$F = \frac{N \text{ CORRECT}}{N} \times 100$$

where N CORRECT is the number of candidates getting the item correct and N is the total number of candidates.

Discrimination values are given by:

$$D = \frac{N_u - N_l}{N/2}$$

where N_u = Score of upper 50%
 N_l = Score of lower 50%

A simple measure of reliability was also calculated - the Kuder Richardson KR21.

$$R \text{ (KR21)} = \frac{K(1 - (M(K - M)/KS^2))}{K - 1}$$

K = number of test items (30)
M = Mean Score (Arithmetic Mean)
S = standard deviation of test scores

KR21 tests if all items measure the same thing (homogeneity) and is roughly equivalent to split-half tests of reliability. It is not appropriate for speeded tests and is not an indication of constancy of score over time.

Analysis of Item Characteristics

A summary of the performance of the test is given below:

	School 1	School 2	School 3	School 4	School 5	School 6
N	64	78	48	41	38	70
Xav	16.56	16.24	14.80	15.07	18.55	20.43
Xbav	17.03	16.05	16.40	15.83	18.36	21.10
Xgav	16.09	16.83	13.20	14.00	18.92	19.72
O	3.07	3.51	3.88	3.05	2.77	3.35
R	0.22	0.41	0.52	0.20	0.08	0.43

Where

N = number of candidates analysed
Xav = average score
Xbav = average for boys
Xgav = average for girls
O = standard deviation
R = KR21 reliability

The analysis for the first five schools was aggregated since these schools were all normal public schools. The analysis of the sixth school, which was residential and an institution with a selected intake, was undertaken separately. Summary results are shown next page.

Item	Schools 1 - 5		School 6	
	Facility	Discrimination	Facility	Discrimination
1	97.1	0.04	98.5	0.03
2	84.6	0.13	100	0.00
3	77.9	0.04	90.0	0.14
4	56.0	0.31	78.5	0.26
5	63.0	0.12	68.6	0.40
6	33.8	0.06	50.0	0.20
7	17.9	0.16	34.3	0.29
8	99.2	0.02	94.3	0.11
9	17.5	-0.03	24.3	0.09
10	15.8	-0.03	21.4	0.20
11	81.7	0.12	87.1	0.09
12	35.0	0.23	54.3	0.51
13	65.0	0.15	80.0	0.29
14	66.7	0.07	75.7	0.20
15	71.7	0.27	74.3	0.34
16	21.0	0.13	34.3	0.06
17	55.6	0.34	78.6	0.37
18	73.8	0.31	85.7	0.29
19	79.7	0.10	88.6	0.23
20	62.9	0.35	60.0	0.51
21	20.9	0.10	41.4	0.43
22	24.0	0.32	77.1	0.34
23	24.5	0.24	50.0	0.60
24	54.0	0.36	81.4	0.31
25	26.9	0.27	48.6	0.51
26	68.5	0.17	60.0	0.34
27	56.8	0.28	67.1	0.49
28	36.1	0.42	61.1	0.26
29	59.4	0.06	97.1	0.06
30	20.7	0.04	41.4	0.43
Average	52.3	0.17	66.8	0.28

From this it can be seen that the average facility in the five schools is at a level we would expect if we wished to maximize discrimination i.e. about 50%. In fact the test discriminates much more effectively in School 6 (0.28 compared to 0.17 on average). If this were a real test then we should be concerned that the items that were designed worked best in separating the performance of more able pupils but not so well with more average performers.

Other data indicate that some items (e.g. 3, 8, 9, 10, 15, 17, 18, 20, 22, 25, 28, 29, 30) were more difficult for girls than boys. It was also true that the data on item performance showed that the power of distractors varied widely from item to item and the reasons for this were examined.

Concluding Remarks

The item writing exercise simulated the creation of a test instrument for Primary Science and Agriculture as part of a training exercise. Participants were unfamiliar with the Kenyan curriculum on which material was based and some had not been directly involved in test construction and analysis previously. The construction of valid and reliable test papers of this kind would be a task that would normally occupy several technically competent staff for some

months. It was therefore impressive that within less than a week we were able to specify, develop, try out, and analyse the performance of a test instrument that we designed.

The primary purpose of this exercise was to develop skills in creating test items. The pre-test that we used had to be modest in scale and the sampling employed could not be representative of the population of pupils in Kenyan schools. Its purpose was to assess our success in creating items rather than reach conclusions about the performance of pupils that might have more general applicability. In this the exercise succeeded.

The analysis drew attention to a number of problems in item construction that included

- items which discriminated poorly or in some cases negatively;
- items where the performance difference between boys and girls was exceptionally large;
- items which were too difficult or too easy to be useful in discriminating performance;
- items where distractors had very uneven power and where some options were so implausible no candidates chose them; items that required a lot of time to understand and complete;
- items that required high quality diagrams to function as intended;
- items which may test language facility more than an intellectual skill.

The results of the trial also provided food for thought in a number of areas:

(1) It was interesting that the instrument designed appeared to function more as expected in School 6 than in the other five schools. In School 6 discrimination was relatively high despite the test being quite easy for most candidates. This implies that most items were understood as intended and the best pupils performed well on each item. In the other schools the picture was more mixed, suggesting that the participants might be better at writing items for the best pupils.

(2) The fact that a number of items appeared easier for boys suggested the possibility that it was the form of the item rather than the intellectual skill that might be responsible for the differentiation. It was not possible to resolve this question without more systematic research. If those items which favoured boys were removed from the test the overall differences in performance would of course reduce. This might or might not be considered desirable depending on what the reasons were for the differential performance on some items rather than others.

(3) The predictions that participants made of the performance of items were fairly inaccurate and on average not much better than guessing. This can be partly explained by their unfamiliarity with Kenyan pupils. It did illustrate the importance of pre-testing items wherever possible to reduce the possibility of poor quality items surviving into national test instruments.

(4) The item construction process resulted in a number of over-complex items being produced that were judged inappropriate for the final instrument. There may be a temptation to design items that are cleverly conceived and impress other item writers but that are too difficult for many candidates.

(5) The importance of analysis of item performance was evident. It helped explain why many items did not function as intended. It suggested that much might be learned by combining examination performance data with teaching and learning studies to establish why differences in performance emerged and how they might be reduced. This was seen as a challenge for the future.

2.4. Bibliography

Black, H.D. & Dockrell, W.B. (1984) Criterion referenced assessment in the classroom. Scottish Council for Research in Education. Edinburgh.

Bloom, B.S., Hastings, J.T. & Madaus, G.F. (1971) Handbook on formative and summative evaluation of student learning. McGraw Hill. New York.

Capper, J. (March 1994) Testing to Learn... Learning to Test. A Policymaker's Guide to Better Educational Testing. Executive Summary. Academy for Educational Development. Washington D.C.

Chimwenje, C. (1993) Primary School Leaving Examinations in Malawi. PhD Thesis, University of Sussex/England.

Colclough, C. with Lewin, K. (1993) Educating All the Children. Strategies for Primary Schooling in the South. Clarendon Press. Oxford.

Dearing, R. (1993) The National Curriculum and its Assessment. Interim Report. Schools Curriculum and Assessment Council. London.

Dore, R.P. (1976) The Diploma Disease. Unwin Education. London.

Ebel, R.L./Frisbie, D.A. (1991) Essentials of Educational Measurement. 5th Edition. Prentice Hall.

Eisner, E. (1968) Instructional and Expressive Objectives. AERA Monographs on Curriculum Evaluation No. 3. Chicago.

Gagne, R.M. (1965) The Conditions of Learning. Holt, Rinehart and Winston. New York.

Harding, J. (July 1992) Breaking the barrier: girls in science education. Unesco: IIEP. Paris.

ILO/JASPA (1981) The Paper Qualification Syndrome and the Unemployment of School Leavers. Contributor to Vol. I (East Africa) and Vol. II (West Africa). International Labour Organisation; Jobs and Skills Programme for Africa. Addis Ababa.

Kellaghan, T./Greaney, V. (1992) Using Examinations to Improve Education. A Study in Fourteen African Countries. World Bank Technical Paper No. 165. The World Bank. Washington D.C.

Keeves, J.P. (1994) National examinations: design, procedure and reporting. Unesco: IIEP. Paris.

Lewin, K. (1992) Science education in developing countries: issues and perspectives for planners. Unesco: IIEP. Paris.

Mager, R.F. (1962) Preparing Instructional Objectives. Fearon Publishers. Palo Alto.

Mauritius Examination Syndicate (Oct. 1992) Learning Competencies For All. Essential and Desirable Learning Competencies for Standards 4, 5 and 6. Towards a holistic approach to examination reform.

Miller, R. (1962) Analysis and Specification of Behaviour for Training. In: Glaser, R.L. (ed.) Training research and education. University of Pittsburgh Press. Pittsburgh, PA.

Njabili, A.F. (1993) Public Examination: A Tool For Curriculum Evaluation. Mtire Publishers. Dar-es-Salaam.

Obura, A.P. (1991) Changing Images. Potrayal of Girls and Women in Kenyan Textbooks.

ACTS Press. Nairobi, Kenya.

Oxenham, J. (ed.) (1984) Education versus Qualification. Unwin Education. London

Pennycook, D. (1990) Factors Influencing the Introduction of Continuous Assessment Systems in Developing Countries. In: Layton, D. (ed.) Innovations in science and technology education. Vol. III. Unesco. Paris, pp. 139-152.

Popham, W.J. (1984) Specifying the Domain of Content Behaviours. In: Berk, R.A. (ed.) A Guide to Criterion-Referenced Test Construction. John Hopkins University Press. Baltimore.

Popham, W.J./Baker, E.L. (1970) Establishing Instructional Goals. Prentice Hall.

Ross, A. (1994) Science Achievement in Papua New Guinea. D Phil. Thesis, University of Sussex, England.

Satterly, D. (1989) Assessments in Schools. 2nd edition. Blackwell. Oxford.

Shephard, L.A. (1991) Will National Tests Improve Student Learning. In: Phi Delta Kappa, November.

Somerset, H.C.A. (1982) Examination Reform: The Kenya Experience. Report prepared for the World Bank, Washington D.C.

Somerset, H.C.A. (1988) Examinations as an Instrument to Improve Pedagogy. In: Heyneman, S.P. & Fägerlind, I. (eds.) University Examinations and Standardized Testing - Principles, Experiences and Policy Options. World Bank Technical Paper No. 78. Washington D.C.

Tsayang, G.T. & Ngwako, A.D. et al. (eds.) (Sept. 1989) Gender and Education. Proceedings of a Workshop. Occasional Paper No. 2. University of Botswana. Gaborone.

Tyler, R.W. (1964) Some persistent questions in defining objectives. In: Lindvall (ed.) Defining Educational Objectives. University of Pittsburgh Press.

Unesco (1993) World Education Report 1993. Unesco Publishing.

Wolf, A. (1993) Assessment Issues and Problems in a Criterion Referenced System. Further Education Unit.

3. Primary School Examinations in Kenya with Special Reference to Item Construction for Science and Agriculture

3.1. The Use of Examination Results for Monitoring Performance of Schools, Districts and Provinces

Francis K. Kyalo, Kenya National Examinations Council

Kenyans wait for the results of national examinations with awe and trepidation. This is so because many people erroneously believe that failing in examinations spells doom. This is attributable to the fact that a person failing in examinations may find it exceedingly difficult either to find employment or to continue with further education or training. Thus, an examination has a tremendous influence on schools, districts and the education system in general. Any information pertaining to examination results is taken with a lot of seriousness by the students, teachers, parents, education administrators and the public at large.

The curriculum which was followed in Kenya between 1963 and 1984 was seen by the »Report of the Presidential Working Party on a Second University in Kenya (1981)« as unadaptable to the changing needs of society.

The Certificate of Primary Education (CPE) examination then consisted of three papers, namely: English, Mathematics and a General Paper consisting of Geography, History & Civics and General Science. To most candidates/school-leavers who did not proceed to the next cycle of education, this general certificate was terminal and they were expected to join the work force. Unfortunately, they had no employable skills.

In the new 8.4.4. system of education, that was examined for the first time in 1985, the primary education curriculum was revamped in structure, duration and content. Previously, candidates had been examined and certified at the completion of seven years of primary education. With the new system of education, the duration was increased to eight years and more emphasis was placed on the acquisition of practical skills. The teaching of pre-vocational subjects such as Art & Craft, Agriculture, Business Education, Music and Home Science was taken more seriously than in the previous primary school curriculum which emphasised the purely academic subjects. Learners are thus being equipped at every level with adequate skills and knowledge for self-employment. It is now much more the responsibility of the candidate to make good use of the knowledge and skills gained through the school system. The certificate is therefore not an automatic ticket to employment.

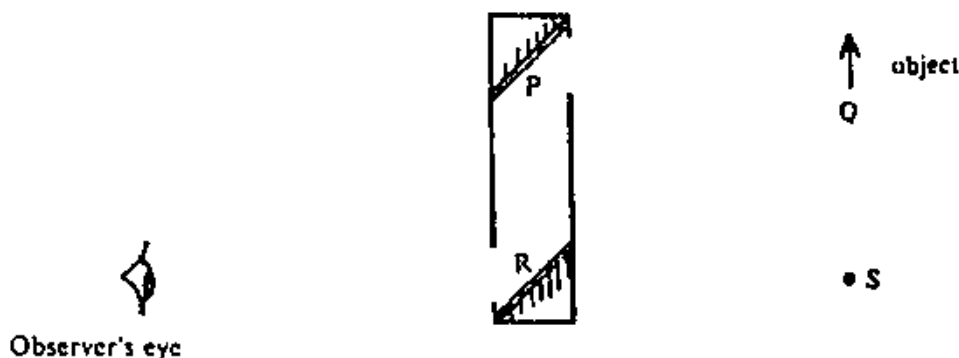
As a matter of policy, it was decided that the students would be tested by means of a national written examination - Kenya Certificate of Primary Education (KCPE) - and assessment of practical work at the school level. This assessment of the practical skills was to be conducted by the Ministry of Education, which was also to issue the results in a revamped school-leaving certificate at the end of the course. The KCPE was to be conducted by the Kenya National Examinations Council.

This called for a change in the examination system. The former Certificate of Primary Education (CPE) was replaced by the Kenya Certificate in Primary Education examination (KCPE) which is tested in seven papers that include, English Language and Composition; Kiswahili Language and Composition; Mathematics; Science & Agriculture; Geography, History & Civics, Religious Education; Art & Craft and Music; Home Science and Business Education.

In the KCPE examination psychomotor skills are tested indirectly. It is intended that questions that are set based on experimental situations test, albeit indirectly, the acquisition, or lack, of the practical skills required for a candidate to respond correctly to such questions.

Question 27, in the 1991 KCPE Science Section, could be cited as an example of questions that are set to test acquisition of practical skills by the candidates.

27. The diagram below represents a periscope, an object and an observer's eye.



The image of the object as seen by the observer will appear at

- A. P
- B. Q
- C. R
- D. S

A periscope uses plain mirrors and therefore the candidates were expected to use their knowledge of the properties of a plain mirror to answer this question. The small percentage of candidates who scored this item correctly shows that most candidates did not have this knowledge. More than half of the candidates in the sample chose option C which shows that the image is formed on the mirror. Anyone who has ever used a plain mirror to view himself/herself knows that this is not true.

One's image always appears to be formed behind the mirror. Simple experiments done by the pupils using plain mirrors would help them to know that the image of an object placed in front of a plain mirror is formed behind the mirror. Thus teachers should seize every possible opportunity to have their pupils perform simple experiments to make their acquisition of scientific concepts easier and faster.

In due course, the Kenya National Examination Council (KNEC) in conjunction with the Ministry of Education will assemble data on the performance of candidates in the national examinations and in the assessment of practical skills at the school level to establish the degree of correlation between the two modes of assessment. This will be done to enhance the quality of instruction at the school level for the acquisition of practical skills.

Before the inception of the KCPE examination, the possibility of the KNEC sending out, to primary schools assessors for the practical component of such subjects as Agriculture, Art & Craft, Music and Home Science was explored. However, this was found to be practically impossible due to the following constraints:

- The present KCPE candidature is quite large (421,617 in 1992 in 12,900 schools) and the logistics of practical tests for such a large number of pupils are overwhelming.
- The short duration in which to administer, mark, process and release results in time for those who qualify to go to secondary schools.
- Lack of materials, apparatus and facilities to conduct practical examinations.
- The large expense that a practical examination would entail.
- The varied conditions in individual primary schools which would make the results of such

an examination inequitable and therefore invalid and unreliable. The results of such an examination, therefore, cannot be used as a basis of certification and ranking of candidates at a national level.

In view of these constraints, practical skills are, in addition to KCPE examinations, assessed at the school level by the teachers and external assessors through projects that are recommended, organized and monitored by the Inspectorate Section of the Ministry of Education. The grades obtained from this school-based assessment are used to award each pupil a School Leaving Certificate, while the scores they obtain in the national examination are used to award each pupil a KCPE Certificate and also to rank them according to their achievement.

The results of national examinations are used to monitor performance of schools and districts. Every year, after the release of KCPE results, the KNEC prepares schools' and districts' order of merit lists. Such order of merit lists are based on the schools' or districts' mean scores. There are two kinds of order of merit lists. One is based on the mean scores on overall performance while the other is based on the performance in individual subjects. Thus, the standing of a school or district in the overall performance, or in each subject, can be determined.

Initially, the order of merit was done for all schools, irrespective of their size. This generated a lot of public outcry because the comparison was regarded as unfair, especially between the large and small schools. The current practice is to compare similar schools in terms of the number of candidates. The order of merit lists enable each school and district to compare its performance against that of other schools and districts. Each school, or district, could also compare its current performance with its previous performance. Thus any improvement, or otherwise, in performance can be noted from the order of merit. Such information has been found to have the following positive influences on the education system:

- The schools and districts that perform very well are motivated to work even harder in order to maintain or even improve further.
- The schools that perform poorly are stimulated to work harder so that they too can be at the top of the order of merit list. In fact, each year a number of schools and districts improve their performance significantly.

This has been made possible by the schools and districts tightening educational guidance, school supervision or even by removing inefficient headteachers, teachers and field staff. In addition, the parents exert pressure and therefore the school authorities are forced to seek for ways and means of improving the performance.

- Ranking of schools and districts stimulates fruitful competition between schools and districts.
- The quality and frequency of school inspection and supervision has improved. In an attempt to improve performance in their regions, the District Inspectors and Assistant Primary School Inspectors are visiting schools more frequently than ever before.
- Teachers in some districts organize mock examinations at the district or divisional levels. In groups, they set and moderate test items to almost the same standard as the Council examinations. The experience so acquired has made the teachers more effective in their instructional duties, in addition to improving their performance.
- Teachers, nowadays, are paying more attention than ever before to the curriculum requirements, its interpretation and implementation. They are more committed to their instructional duties. They teach before and after the official working hours.
- It has encouraged the school committees and the parents'/teachers' associations in schools to be more interested, not only in the acquisition of physical facilities, but also in

the quality of instructional programmes of the schools. In some schools, for instance, the teachers are paid a honorarium for working outside their normal working time.

Some negative influences of the ranking of schools and districts have been cited by educators and members of the public. One such influence is where teachers force pupils to repeat some classes. This has already been rectified by the Ministry of Education by issuing a warning that such a practice should stop. This has borne results because the increase of candidates in 1992 (421,167) over last year's (362,093) is quite significant. This also shows that the rate of repetition is going down. Another negative influence has been the overworking of the primary school pupils by teachers who give instruction outside the official working time. This practice has been a result of overloaded syllabi. The problem has, however, been addressed during the recent review of the curricula for the 8.4.4 education system. The syllabi have either been pruned or re-organized in 1992 to facilitate effective coverage within the time allocated for the teaching of each of the subjects.

Similarly, after the release of the KCPE examination results, the Council prepares a KCPE Newsletter. In this Newsletter, comments on the candidates' performance in individual questions, and papers as a whole, are provided. The candidates' weaknesses, what was expected of them and suggestions for improvement in the teaching and learning procedures are highlighted. Corrective measures are also recommended where the teachers and pupils seem to misinterpret the official curriculum. Thus examination results are used to monitor how the schools are implementing the curriculum in the process of curriculum appraisal. Furthermore, the teachers are given an opportunity to suggest more effective instructional methods to the KNEC for dissemination to other schools. This is done by means of a questionnaire at the end of each issue of the KCPE Newsletter.

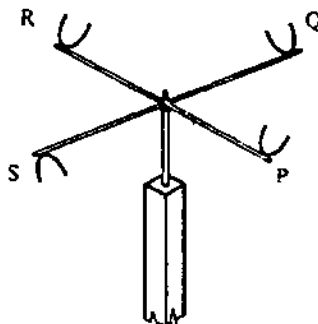
National examinations in Kenya generate a lot of data on candidates' performance. Some of this data reflects the manner in which the requirements of the curriculum are interpreted by the teachers. For instance, the KCPE Science syllabus may require students to be involved in the learning/teaching situation by being allowed to carry out many simple experiments, make and record their observations, and draw conclusions from these observations. Well-designed Science examination items can be developed to test whether or not this requirement has been met by the teachers.

The KNEC uses the data generated by item analysis to highlight incorrect interpretation of the official curriculum by the teachers and pupils alike. The data so generated is then used to suggest corrective measures to be taken by those charged with the responsibility of developing and implementing the curriculum.

A KCPE item, such as the one cited below, clearly shows how it is possible to set an item which tests how well the candidates have acquired the practical skills as per the syllabus requirements.

Question 9 (1991 KCPE Science Section)

A pupil constructed a simple anemometer as shown in the diagram below.



Which one of the cups should be correctly placed so that the instrument can function properly?

- A. P
- B. Q
- C. R
- D. S

The question was testing the candidates' understanding of the construction and use of a simple anemometer. Candidates who answered the question correctly were above average in their performance in the Science section of the paper. The fact that about half of the candidates in the sample chose option D, which is incorrect, suggests that most pupils had never constructed or used an anemometer. In most cases, its use may have been illustrated to them by means of a drawing on the chalkboard and therefore the pupils found it difficult to understand. If pupils were given a chance to construct and use an anemometer, very few would have problems with such a question.

Thus, well-designed test items assist in strengthening the process of curriculum implementation and appraisal.

The foregoing discussion suggests very clearly that examinations have a very positive role to play in educational development. In order to maximize their contribution to educational development, it is necessary that examinations reinforce the objectives for which they are meant. They can do this by testing the knowledge and skills that are appropriate and relevant to the majority of the candidates. Finally, examinations should also assist in the process of curriculum implementation and appraisal.

3.2. Testing and Monitoring Procedures Developed for Primary Schools

Paul M. Wasanga, Kenya National Examinations Council

The Purpose of Testing and Monitoring Pupils

When trying to establish the worth of anything, and hence evaluate it, we need information and yardsticks against which to judge not only the information we require, but the information we receive. In education we are basically concerned with the worth of such things as curricula, teaching methods and course materials. One major significant source of information, although not the only one, is the performance of those being taught -»the pupils«. Since performance can be determined by testing, one needs to concern oneself with the question, why test (assess) and monitor pupils?

Below are some reasons that have been advanced for assessment and monitoring of pupils. It is in order to:

- gather information about a wide range of the pupils' characteristics as a feedback for making decisions;
- accumulate records of progress for the pupils;
- provide information from which teachers can obtain insights into their own effectiveness in teaching;
- allocate pupils to sets or groups;
- compare new teaching materials with old ones;
- give incentive to learning and aid to remembering;

- determine pupils' strengths and weaknesses;
- predict pupils' future performance;
- determine the effectiveness of the methods of instruction;
- determine the suitable candidates for further education in institutions of higher learning (e.g. secondary schools);
- assign individual pupils grades that show how different their abilities are from those of other pupils.

To be able to carry out all the above tasks it is important to develop measurement procedures that can fulfil such requirements well. The procedures used to assess and monitor the primary school pupils in Kenya are discussed below.

Methods Developed for Testing and Monitoring Primary School Pupils in Kenya

The primary school pupils in Kenya are assessed using two methods, school-based continuous assessment and Kenya Certificate of Primary Education (KCPE) examination offered by the Kenya National Examinations Council.

School-based Continuous Assessment

Continuous assessment involves updating of judgements about performance of the pupils. It therefore should be:

- systematic;
- objective;
- comprehensive;
- cumulative and
- guidance-oriented.

Many educators hold the view that continuous assessment of the pupils is best carried out by the class teacher. This, of course, is based on the assumption that the teacher is almost always in close contact with the student and his/her assessment should therefore provide a reliable and valid assessment of the pupils' performance. Since such an assessment is done in schools by the teachers, it is called school-based continuous assessment.

Instruments Used for School-based Assessment

To determine the extent of the pupils' learning, various instruments are used to measure the pupils' performance in Kenyan primary schools. Some of the commonly used instruments are:

- Exercises

These are questions, assignments, quizzes etc. set and administered by class teachers to pupils at regular intervals or at the end of a specific topic. Pupils are expected to complete the exercise during class-time/period and hand over the work to the teacher. The main aim of these exercises is to determine whether the pupils have mastered/learned the topic covered. Thus, most of such exercises normally test recall of knowledge.

- Terminal tests

These are tests that are administered by the teacher/school at the end of each term. Such tests cover topics/subject matter taught during the term. These tests are done under controlled time limits and similar conditions. Such tests normally sample all major topics covered in a term. These tests are more valid than the exercises because, although the teachers who set them may not have experience in setting tests that measure all the skills, the tests do not test recall of facts only.

- Home-work

This is another method used to test and monitor pupils' progress in Kenyan primary schools. In this method the pupils are assigned work at the end of a lesson. The pupils then take it home. There is no time restriction and the pupils are even free to consult books, friends or even parents. These home-work exercises reinforce teaching and enable the pupils to identify areas not clearly understood during class period.

- Projects

In addition to the final examination (KCPE) the primary school pupils are also assessed by their teachers through projects recommended by the inspectorate of the Ministry of Education. In the Kenyan context, a project may be defined as any set task from which time constraints have been largely removed.

The projects offered to the primary school pupils are normally chosen from the courses which are supposed to have been covered in Standard 7 and 8; for example in Art and Craft, sculpture, containers, body covers, fabric decoration, graphic design, wood work, metal work, leather work etc. may be chosen as projects, while in Music pupils may be asked to make musical instruments. All the other practically oriented subjects like Science and Home Science are organized along similar lines.

Marks obtained from such course work (projects) are normally incorporated in the pupils' leaving certificates.

- Zonal/district organized tests

These are tests that are set and administered by teachers in zones/divisions or even in districts. These tests might also be referred to as inter-school tests. To set these tests, a number of experienced teachers is selected from different schools in the particular zone. These teachers set the tests, and selected officials organize the administration and marking of such tests. Zonal tests are usually given to primary pupils at the end of class 7. Marks obtained from such tests are used to predict the performance of the pupils in the Kenya Certificate of Primary Education examination.

- Fieldwork

Fieldwork involves a visit to and the study of an area outside the classroom by the pupils. Fieldwork calls for the pupils' skill in observation and mastery of concepts and accurate recordings. Before the fieldwork starts, the teacher normally identifies the area of study and the task the pupils are expected to carry out. The teacher then visits the site, and arranges for transport if any is required. During fieldwork pupils are expected to discuss their observations and record them. The teacher then may ask the pupils to write a report. The teacher gives marks to the pupils based on their recordings and reports.

Use of the Information Obtained from School-based Continuous Assessment

The information obtained from school-based assessment is usually kept by the teachers and also by the school administration. Such information is used for:

- gauging the teachers' effectiveness in teaching;
- guiding teaching;
- providing motivation to the pupils;
- providing feedback information to the parents and other interested people to aid counselling and decision-making;

- comparing the pupils with others from different schools and suggesting areas of improvement;
- continuity of performance of the pupil.

This information should therefore be stored in such a way that it is:

- permanent;
- easily understood;
- easily interpreted;
- easily retrieved.

Problems Associated with School-based Continuous Assessment

These include:

- inflation of scores;
- fake scores;
- missing scores;
- absent pupils;
- testing of recall of facts by the teachers because of lack of knowledge on measurement and time for writing questions that test higher abilities;
- unreliability of scores;
- subjectivity of teachers;
- unavailability of trained teachers;
- lack of uniformity in facilities available in primary schools across the country.

Because no system has been developed to eliminate such problems, in Kenya continuous assessment scores are not included in the Kenya National Examinations Council certificate for primary school leavers.

Kenya Certificate of Primary Education Examination

At the end of the eight years of primary schooling the Kenya National Examinations Council offers a written examination. This examination is called the Kenya Certificate of Primary Education (KCPE). This examination replaced the unpopular Certificate of Primary Education (CPE) which has been accused of being biased towards recall questions only and failing to adequately examine practical skills.

The Kenya Certificate of Primary Education examination offers a broader concept of assessment by testing many subjects and more complex cognitive processes like reasoning. This is demonstrated by the questions that are designed to test practical skills. These questions are set in such a way that they are able to sort out those who have been exposed to practical skills and those who have not. This is especially so for papers like Science and Agriculture, Art, Craft and Music, Home Science and Business Education.

To come up with questions that test practical skills and other higher abilities, using multiple-choice format is not a simple matter. The Kenya National Examinations Council, however, uses experts and experienced examiners to develop papers that meet this challenge. This, of course, involves hard work and dedication on the part of those who set and moderate such questions. A discussion of what goes on during setting and moderation of Science questions is illustrated below.

Setting of Science Questions at KCPE Level

Science questions are set by a group of experts who have a good knowledge of the primary school Science curriculum. These experts are drawn from the whole country. Normally, ten experts are involved in this exercise. The first exercise during the setting involves drawing a table of specification.

To draw such a table, the experts have not only to be conversant with the objectives of the curriculum but also have to carefully re-study them. This is done so as to make sure that all the content in the syllabus is tested and also that all the skills are tested. Below is an example of a table of specification drawn by setters and moderators for Primary Science KCPE examination.

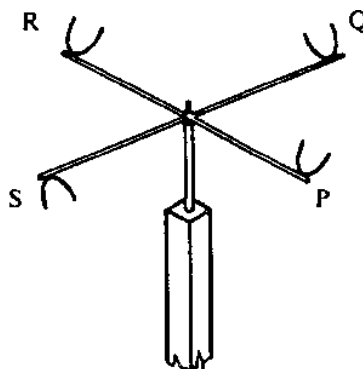
Table of Specification for Science KCPE						
	Abilities	Knowledge	Comprehension	Application	Higher Abilities	Total
Content Area						
1. Energy		1	1	1		3
2. Living Things		3	2	1	2	8
3. Properties of Matter		1	1	2	2	6
4. Environment			1		1	2
5. Making Work Easier			1	3	1	5
6. Weather and Astronomy			2			2
7. Soil		1	1			2
8. Health Education		1	1			2
Total		7	10	7	6	30
		23.33%	33.33%	23.33%	20%	

This table of specification shows that the content in primary science can be summarized in eight broad areas. According to the objectives in the syllabus the skills are weighted. Using this table, questions are set in such a way that:

- 23.33% of them (7 questions) test knowledge (recall of knowledge);
- 33.33% comprehension (10 questions);
- 23.33% application (7 questions);
- 20% higher abilities (6 questions).

Thus, a paper set using this table does not test recall of facts only but also all other skills demanded by the curriculum. Below are some questions that have been used before to test skills other than knowledge.

(1) Question 9, KCPE Science 1991:



A pupil constructed a simple anemometer as shown in the diagram.

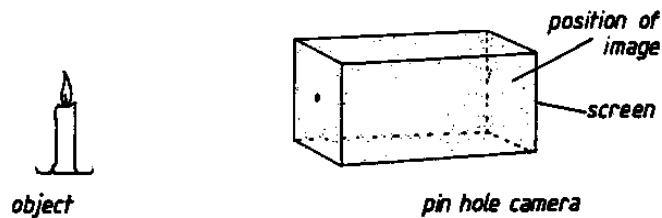
Which one of the cups should be correctly placed so that the instrument can function properly?

- A. P
- B. Q
- C. R
- D. S

This question requires the candidate first to have knowledge of what an anemometer is (an instrument used for measuring the force/speed of wind), and then look at the diagram and work out which cup is placed wrongly. This question cannot be said to test recall only; it is also testing a practical skill because the candidate has to work out how the cups should be placed for the instrument to function properly.

(2) Question 12, KCPE Science 1988:

Study the arrangement shown in the diagram below and answer the question that follows.



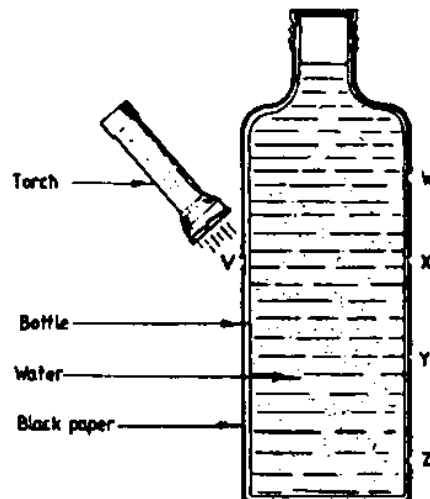
Which one of the following statements is correct? The image formed on the screen is

- A. upside down
- B. brighter than the object
- C. round in shape
- D. bigger than the object.

The question requires the candidate to construct lines from the object through the pin-hole to the screen so as to determine how the image would be. Obviously this cannot be recall.

Although the question itself is a multiple-choice question the skill tested here is a practical skill which involves reasoning. The two questions below serve as further examples of such questions.

(3) Question 4, KCPE Science 1990:

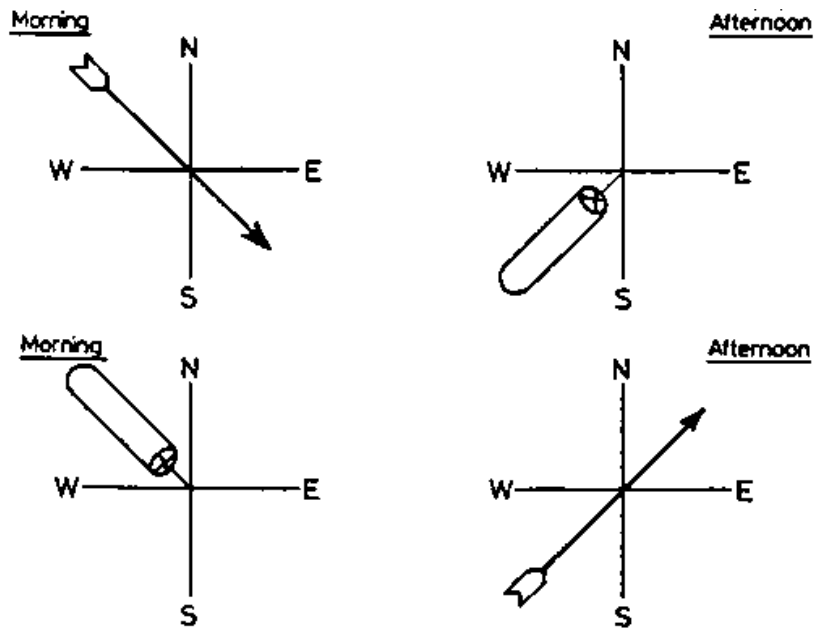


Kamau filled a transparent bottle with water and then covered it completely with black paper and made a small hole, **V**, on one side and four other small holes **W, X, Y, Z** on the opposite side. He shone light from a torch into the water as shown in the diagram.

At which point is the light most likely to shine out?

- A. **W** B. **X** C. **Y** D. **Z**

(4) Question 16, KCPE Science 1989:



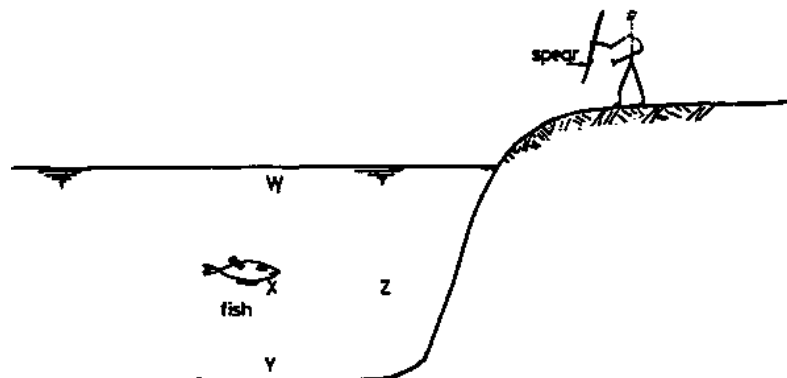
Omar made a wind vane and a wind sock and placed them in an open place. He then observed them in the morning and in the afternoon. The positions of the instruments were as shown in the diagram below.

Which one of the following gives the correct record of the wind direction for the day?

- | | Morning | Afternoon |
|----|----------------|------------------|
| A. | SE | SW |
| B. | NW | NE |
| C. | NW | SW |
| D. | SE | NE |

(5) Question 12, KCPE Science 1989:

A fisherman looks down into a lake and sees a stationary fish as shown in the diagram below.



At what point should he aim his spear in order to have the **BEST** chance of hitting the fish?

- A. **W**
- B. **X**
- C. **Y**
- D. **Z**

Questions testing higher abilities than recall and practical skills are not confined to Science only. Such questions are also found in all the papers offered in the Kenya Certificate of Primary Education examination. Practical skills are, however, commonly tested in Art, Craft, Music, Agriculture, Home Science and Business Education as illustrated by the following questions.

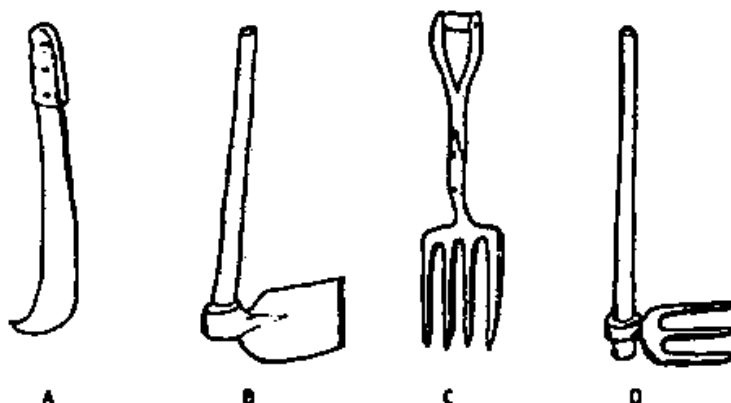
(6) Question 19, KCPE Art and Craft 1988:



Which one of the shadows of the figure is correct if the sun is at the position as shown in the illustration?

(7) Question 52, KCPE Agriculture 1990:

Which one of the tools illustrated in the diagram below is suitable for digging a hard ground?



All the questions given as examples above clearly indicate that the Kenya National Examinations Council has made great strides towards the production of paper and pencil examinations that are capable of testing higher abilities and practical skills instead of recall of straightforward knowledge at primary school level. It should, however, be understood that the Kenya National Examinations Council examines and grades Std. 8 pupils from the results obtained in the Kenya Certificate of Primary Education only. The Kenya National

Examinations Council does not offer continuous assessment tests mainly because the KCPE candidature is very large (about 421,617 at present) and owing to the problems associated with continuous assessment discussed earlier.

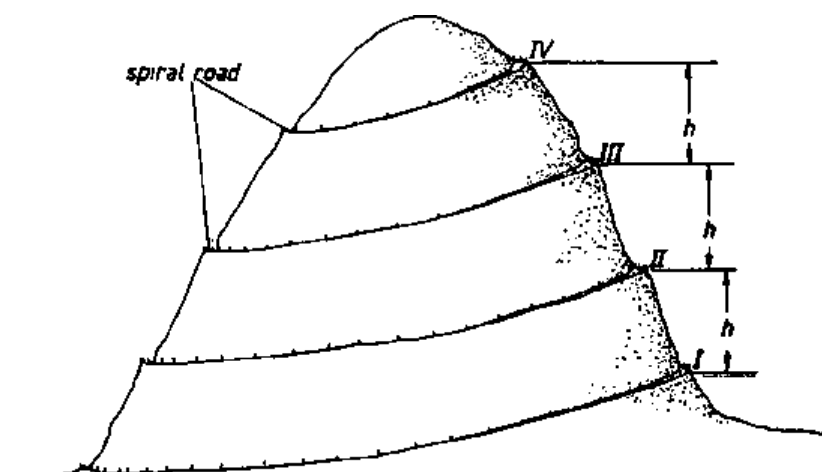
Use of KCPE Results for Monitoring Primary School Pupils

Unlike school-based continuous assessment, which is used continuously to monitor pupils' progress in education, KCPE results are only used to advise the teachers on areas of weaknesses so that they may re-assess their strategies of teaching. This advice is normally given in the form of backwash documents written by the Council. This document is called the **KCPE Newsletter**.

The KCPE Newsletter discusses questions done poorly by the Std. 8 candidates and suggests how misconceptions, that lead to the poor performance in the particular questions, can be corrected. The two examples below serve to illustrate this point.

(1) Question 20, KCPE Science 1989:

Naposho walked to the top of the hill along the spiral road represented by the diagram below.



If the heights, h between the points are all the same, the effort used during the climbing was

- A. least between points II and III
- B. least between points III and IV
- C. least between points I and II
- D. the same between all the points

Response Pattern

Option	A	B	C*	D
% Choosing the Option	6.27%	33.97%	28.73%	30.55%
Mean Mark on other Questions	26.59	31.7.1	31.83	30.45

Question 20 was testing application knowledge in a new situation. It was scored correctly by only 28.73% of the candidates who were slightly of above average ability (had a mean mark of 31.38 compared to the sample mean of 21.01). Options B, C and D were chosen by candidates of almost similar ability because their means were less than a mark about the sample mean. This indicates that most candidates were only guessing the answer and hence the very small Discrimination Index (0.01). It is possible that most candidates failed to see the relationship between what they had learned in class and what was being tested in this question. Teachers are advised, therefore, to relate what pupils learn in class to real-life situations outside the classroom, for this is the ultimate goal in education. Knowledge that cannot be applied to solve problems and help us understand the world around us is dormant and hence worthless.

(2) Question 19, KCPE Art and Craft 1988:



Which one of the shadows of the figure is correct if the sun is at the position as shown in the illustration?

Choice of the wrong options A, C and D in the question clearly indicates that the candidates may not have had practical exposure of body figures in the sun to enable them determine the sizes of shadows at various positions of the sun.

Conclusion

This paper outlined several ways in which the Kenyan primary school pupils are tested and monitored. School-based continuous assessment was looked upon as a systematic, objective and comprehensive way of collecting information about pupils. Exercises, terminal tests, home-work, projects, zonal/district organized tests and field-work were discussed as some of the instruments of doing this. The Kenya Certificate of Primary Education examination was also discussed as a procedure of testing and monitoring primary school pupils. It was demonstrated that questions offered at KCPE level do cover a variety of skills including practical skills.

3.3. Developing Tests for Complex Cognitive Processes

Obadiah Mucheru, Kenya National Examinations Council

Introduction

Most school curricula are designed with the explicit purpose of inculcating cognitive, psychomotor and affective skills in students. The success of any curriculum in fulfilling this objective can be assessed through the administration of tests at any stage of its implementation. Hence one of the major purposes of assessment in education is to investigate the extent to which students have gained from instruction.

The success of such an investigation would be enhanced if:

- the stock of knowledge possessed by students at the beginning of a programme of instruction is known;

- the objective of instruction or the structure of knowledge the students are expected to acquire is clearly stated and understood;
- the investigator can develop relevant testing instruments that can be used to show the gains made by students at any stage of instruction.

In the following we discuss the efforts item writers should make to ensure that they write items that test higher order thinking abilities.

Instructional Objectives

Instructional objectives are statements of what the instruction is expected to accomplish. Instructional objectives should be expressed in terms of quantifiable, measurable outcomes. Ebel (1963) distinguishes between the explicitly and implicitly stated instructional objectives and goes on to state that objectives that have been prepared to guide instructional planning or to communicate intended learning outcomes to students can also be used for evaluation planning and test development.

According to Ebel, an explicitly stated objective contains a verb that indicates, in operational, behavioural, or observable terms what the learner must do to demonstrate attainment of the objective. He gives the following examples of verbs that distinguish explicit and implicit statements of instructional objectives.

Verbs that Distinguish Explicit Statements of Instructional Objectives

Explicit, Behavioural	Implicit, Non Behavioural
Observable	Inferential
identify, explain, describe, summarize, select, develop, predict, differentiate, define, compare, write	know, consider, understand, discuss, realize, remember, judge, perceive, think about, comprehend, imagine

While developing tests the developer should translate the knowledge and abilities students are expected to acquire into tasks (test items) that require a demonstration of the students' achievements.

Taxonomy of Educational Achievements

A number of educators have developed taxonomies of educational outcomes for use by test developers. Three such classifications are shown in the figure below:

	Bloom's Taxonomy	Ebel's Relevance Guide	Gagne's Learning Outcomes
A.	Knowledge	Terminology Factual Information	Verbal Information
B.	Comprehension	Explanation	Intellectual Skills Cognitive Strategies
C.	Application	Calculation Prediction	
D.	Analysis		
E.	Synthesis		
F.	Evaluation	Recommended Action; Evaluation	
G.			Attitudes
H.			Motor Skills

The taxonomies classify all educational objectives into a hierarchy of categories based on presumed complexity (Gray, 1991). Each succeeding category involves behaviour believed to be more complex than the one previous and each is considered to be prerequisite to the next.

Cognitive Abilities

The term cognition refers to an act or process of knowing that involves the processing of sensory information i.e. perception, awareness and judgement. Cognitive learning includes problem solving, observation, concept formation and creative thinking. In more practical terms students who have followed a curriculum whose aim is to cultivate a wide range of cognitive skills should be expected to demonstrate some or all of the following abilities:

- (1) Ability to identify terms and definitions.
- (2) Ability to identify the reasons and conclusions of a piece of reasoning.
- (3) Ability to judge whether an argument is sound.
- (4) Ability to predict an outcome.
- (5) Ability to explain or illustrate an event, principle etc.
- (6) Ability to recommend appropriate action to solve a problem.
- (7) Ability to solve specific mathematical problems.

The abilities numbered (2) - (7) are fairly complex in nature and any tests designed to elicit the extent of students' achievement of the abilities are likely to be complex.

However, most test developers are more acquainted with Bloom's taxonomy of the cognitive domain. In using Bloom's taxonomy it can be safely assumed that any item that tests beyond knowledge tests higher (complex) thinking abilities.¹

¹ See 2.2.3. Table 2: Major Categories of Cognitive Objectives with Examples

In writing items the tasks should be stated in a style that will lead the student to engage in the desired thought processes.

Types of Questions for Different Cognitive Abilities

Knowledge - The pupil is expected to recall learnt information.

- name the - - -
- label the - - -
- state the procedure used to - - -
- define the term - - -
- describe how - - -

Comprehension - The pupil is required to show an understanding of the general nature of any new material or information.

- state the reasons for - - -
- indicate the advantages and disadvantages of - - -
- why is material/tool Q used for - - -?
- what does the phrase/statement/formular etc. tell us about - - -?
- state in words what the graph/picture etc. shows about - - - .

Application - The pupil is expected to apply what has been learnt to a problem which is new to him/her.

- what advice would you give to e.g. a farmer, a painter etc. to solve the problem of - - -?
- what is likely to result from - - - (unfamiliar set of circumstances)?

- use a graph to show that - - -
- solve - - - (a numerical problem).

NB: It is important to note that the problem whose solution is sought should be new to the target group, and not to the writer of the item. When we forget this cardinal rule, we often fail to recognize such questions as testing complex skills.

Analysis - The pupil is expected to be able to break down materials into its component parts and identify its organisation, structure and the relationship between parts etc.

- name the separate parts of - - - (e.g. a plant, a given instrument etc.)
- show in the form of a diagram the component parts of (a given instrument, object, tool etc.)
- identify the inconsistencies that are apparent in (a given piece of evidence or between two or more pieces of evidence).

Synthesis - The pupil is expected to put together pieces of evidence, or information to produce the desired results e.g.

- the pupil should be able to use an organizing principle (temporal, behavioral, causal) to hold material together.

Evaluation - The pupil is supposed to judge the value of a given piece of information.

- is the evidence given adequate to enable one to make the conclusion that - - -?
- how reliable is the evidence available - - -?
- to what extent is (a given phenomenon) influenced by - - -?

Summary of General Rules for Setting Higher Order Items

- Avoid questions that start with words such as who? what? when? where?. These tend to solicit factual information.
- Use terms like **why?** - - **because** - - **which of the following statements** - - etc.
- Give a set of conditions and ask students to predict a future result.
- State a problem and ask students to suggest a solution.
- Always think of problems to be solved rather than facts to be remembered.
- Always strive to use real life problems - not abstract concepts.
- Avoid using long winded statements.
- Use as much stimulus materials as possible. These include diagrams, graphs, tables, pictures, etc. They are easier to write items on.

The Specifications Grid

The first task of a test developer is to draw a specifications grid to show the topics and skills to be tested and the number of items to be developed in each case. While most item writers use Bloom's taxonomy of the cognitive domain to indicate the skills to be tested other taxonomies may be used. In Science subjects it is not unusual to use science skill processes such as observation, measurement, recording, predicting, etc. instead of Bloom's taxonomy. The most important thing is to have a plan for the test to be developed.

Moderation

The moderation of items should be done by a panel of subject and measurement experts. They should scrutinize each item and improve it or discard it depending on its relevance to the desired goal. The moderating panel must ensure that:

- every item has face validity, i.e. each item appears to test what it purports to test;
- the intended key is the single correct or the best choice among the given responses;
- all the options are plausible and that the distractors are indisputably wrong;
- each item is set at a reasonable level of difficulty for the average candidate;
- the phrasing of each item is unambiguous;
- each item is independent i.e. picking the right answer in one item is not dependent on the student's ability to pick a correct answer in another item in the paper;
- the item does not contain irrelevant difficulty - e.g. long statements which can be shortened without changing the intended meaning;
- the item does not test trivial knowledge;
- the item is free from bias i.e. that the item neither gives an advantage nor a disadvantage to particular candidates on the basis of factors which are extraneous to the purpose of the test;
- each item correlates to the whole test;
- all diagrams are accurately drawn and labelled.

An item which does not meet any of the above criteria should either be improved or substituted.

Evaluation of Test Items for Bias

The evaluation of test items for bias is an important step in test development. The evaluation should be done by a person or persons who are truly conversant with the target population for whom the test is designed. It can be done at the moderation stage or after the moderation. The evaluation for bias can be done by asking a series of questions such as the following:

- Is the language used in the test items, as well as the activities reflected in the items, likely to be offensive to members of any group? i.e. Are the items culture free?
- Are the activities reflected in test items relevant to the life experiences of the persons responding to the items? e.g. Do the items have a rural or an urban bias?
- Are the connotations of key words and phrases in the test items essentially the same for all students?
- Are the test items written in a straightforward, uncomplicated, easily read manner?
- Is the content of test items so interesting to examinees that they will be distracted from the task at hand?

Arrangement of Items in the Test Paper

Most students are filled with anxiety when they are confronted with tests. This anxiety may boil over if the first few items in a test paper require the student to engage in complex thought

processes. As a general rule a test paper should start with easier questions, the more difficult questions appearing towards the end.

Observations and Conclusions

- The term complex-cognitive skills can be used to refer to the skills that go beyond the knowledge or recall level.
- It is difficult for any two people or any two groups of experts to all the time agree on the skill a particular item is testing. The mental processes that the student has to engage in is the single most important determinant of the skill the item is testing.
- When items are meant to reveal whether the student can apply knowledge to solve new problems, care must be exercised to ensure that the problems are new to the student, not to the item writer.
- It is important for the item writers at any level to be fully conversant with what goes on in the classroom. They should also have a good mastery of the subject matter and the objectives of both the curriculum and the particular subject for which they are expected to produce items.
- The task in every test item must be clearly stated. The stem should be stated in the form of a question whenever possible.
- It is easier to test the higher abilities using the »best answer« rather than the »absolute correct answer« type of items.
- The moderation of a test paper should be regarded as a most important stage in test development. Moderators must be fully conversant with the requirements for education in the country. They should also be subject and measurement experts. And since elimination of bias is part of the moderation exercise, the moderators should be fully conversant with the target population.

References

Ebel, R.L. (1963) The Social Consequences of Educational Testing. In: Anastasi, A. (Ed.) Testing Problems in Perspective. American Council of Education.

Gray, L.R. (1991) Education Evaluation and Measurement Competences for Analysis and Application. 2nd Edition. Macmillan Publishing Company.

3.4. Skills in the Construction of Science Tests

Grace Kigoto, Kenya Institute of Education

Introduction

Constructing a good test is not easy, yet the curricular changes witnessed in many of our African countries, especially in the last decade, have demanded more rigorous testing. At the same time, sharp criticisms and bitter attacks have been raised against testing. But it is a well known fact that testing is an essential part of teaching and learning. Efforts have, therefore, to be made by one and all to acquire knowledge, skills and attitudes necessary for the construction of good tests.

Requirements for Constructing a Good Test

A good test can serve several purposes. For example:

- it can provide teachers with relevant information for improving instruction;

- it can help pupils improve their learning through motivation;
- it can facilitate educational planning and decision making.

The Kenya Certificate of Primary Education (KCPE) examination offered at the end of the primary cycle (8 years) is intended to guide in:

- ranking candidates according to their achievement in subjects and therefore selection to secondary schools;
- awarding certificates;
- data collection that can be used by teachers, curriculum developers and educational planners.

An item writer should possess an understanding of the following:

- the structure and goals of the curriculum to be tested;
- the behaviours and processes to be learned;
- the psychological and educational characteristics of the learner.

A good test must be carefully planned in order to serve the purpose it is intended for. The following procedures have been suggested:

- State the general objective of the course and define each objective of the unit in behavioural terms. (See Appendix I for action verbs to use in stating objectives behaviourally and Appendix II for Science Process Skills and Attitudes).
- Make an outline of the content to be covered or that already covered during the course.
- Prepare a table of specifications.
- Construct test items to measure the students' behaviour specified in the table of specifications.

Specifying Objectives to Be Tested

This indicates the learning outcomes that the test developer will accept as evidence that the lesson objectives have been achieved. Secondly, this specifies the student behaviour that is to be measured by test items.

Bloom's taxonomy of intellectual abilities and skills provides categories as follows:

<u>Knowledge:</u>	This provides mainly recall of knowledge e.g. terms, facts, rules and principles.
<u>Comprehension:</u>	This involves understanding of what is being communicated, so that the individual can make use of the material or idea.
<u>Application:</u>	This concerns the use of data, in particular to handle general ideas, rules of procedure or generalised methods, as demonstrated within the category.
<u>Analysis:</u>	The breakdown of information into its constituent elements or parts so that its ideas are made clear.
<u>Synthesis:</u>	Involves putting together elements and parts to form a whole. An individual can take a variety of data and arrange it in such a way as to constitute a pattern.
<u>Evaluation:</u>	This involves the ability to exercise judgement.

Outlining the Content

After specifying the objectives to be tested, the tests constructor should make an outline of the content to be covered by the test. This ensures adequate sampling of the subject matter

of the course. This content outline should show specific elements of the content.

Preparing the Table of Specifications

This is the most crucial stage in the construction of a test, some times called a test blue-print. A table of specifications ensures balance and comprehensiveness of a test. The table has two dimensions. The first represents the different abilities that the student should display and the second represents the subject matter.

Table of Specifications: Science and Agriculture					
Skill	Knowledge	Comprehension	Application	Higher Abilities (Analysis, Synthesis, Evaluation)	Total
Content					
Health Education: Diseases & how to prevent them; Cleanliness	1	1	1	-	3
Energy: Sources, forms, uses and conservation	1	1	1	1	4
Making Work Easier: Construction	-	1	1	-	2
Properties & Characteristics of Soil: Sedimentation, water retention, stickiness and use of locally available soil	1	1	1	-	3
Soil Fertility: Fertilizers, manures, mulching and soil erosion	1	1	1	1	4
Pollution: Pollutants and the environment	1	1	1	-	3
Land Utilization: Uses and farming systems	1	1	1	-	3
Farm Economics: Effect of spacing and fertilizer application, farm records and simple accounts	1	-	1	1	3
Total	7	7	8	3	25

Each row indicates the number of items to be used. It is important that the number of items on each area of content and to each objective in the table of specifications reflect the relative emphasis given during instruction.

Construction of Test Items

Depending on the purpose of the test and the mental abilities students should display, the test constructor will use different types of tests.

Free Response Type

In these, the student is expected to create rather than to choose a correct response. There are completion items and essay types. In completion items the student gives one or two words or a phrase to complete a sentence.

Essay type of question items are easy to construct and when done properly, they can measure complex learning outcomes. They allow for creativity and originality. However, there are disadvantages of using these tests. For example, they can be highly unreliable. It is also difficult to cover all topics using essay type of questions.

To be clear and specific, such action verbs like »describe«, »identify«, »list«, »compare«, »distinguish«, »give reasons for...« should be used when constructing essay questions. To increase reliability, the evaluator should prepare a marking scheme showing the main points expected and the distribution of marks.

Choice Items

These can be marked objectively and are easy to score. They are, however, difficult to construct and can be open to guessing. There are 3 types, namely, multiple-choice items, matching items and true-false. We shall only discuss multiple-choice items in this context.

A multiple-choice item has two parts: a stem and options. The stem can be written as a question, or an incomplete statement. Of the options, one is the answer or **key** while the other options are called **distractors**.

The following should be borne in mind when constructing multiple-choice items:

- The stem should be clear and should include a single problem only.
- Avoid placing the correct answer too frequently in any one position.
- Avoid having similar or related words in both stem and correct answer.
- Check that no item has more than one key.
- Avoid making the correct answer longer or shorter than the distractor.
- Avoid using choices that are unrelated to the subject of the item or are obviously wrong.
- All choices should be plausible.
- Each response should be preceded by an identifying number.
- As far as possible, avoid the use of negatives.
- Each item should stand on its own and no item should provide a clue to the answer of another item.
- The distractors must be a result of common errors or a misunderstanding. (See Appendix III - Test Construction Guidelines)

The above are steps/hints towards the construction of a good test. Before a test is finally declared good, the following needs to be done:

Content Analysis

The content validity is assessed. It is carried out to determine the extent to which test items adequately measure the objectives of the course.

Item Analysis

This takes place after the test has been administered to the learners. Item analysis helps one to identify the objectives and content that has been mastered by the learners and those that require further training. Two types are discussed here:

(1) Item Difficulty

This is calculated by getting the number of learners who got the item right out of the total number of learners who did the test e.g. if 16 out of 20 learners got the item correct, then we have $16/20 = 0.8$. The higher the value of item difficulty, the easier the item is. This shows that most of the learners answered the item correctly and so had achieved the objective.

Another method of calculating item difficulty is by using the formula

$$P = \frac{C}{T} (100)$$

Where P is the percentage of pupils who answered the test item correctly, C is the number of pupils who answered the test item correctly, and T is the total number of pupils who attempted the test item.

E.g. suppose that an item is answered correctly by 40 students out of a total of 60 who attempted it. The level of difficulty P will be

$$P = \frac{40}{60} (100) = 66.6\%$$

(2) Item Discrimination

This shows the extent to which a test item distinguishes between bright learners and weak learners. The following method is used:

- arrange all marked answer papers of the learners in ascending order starting with the paper with lowest mark at the bottom;
- calculate the total number of learners at the top one-third and the bottom one-third. The middle one-third is NOT used;
- count the number of learners in the high score group who answered the item correctly. Call this H;
- count the number of learners in the low score group who answered the item correctly. Call this L;
- calculate the item discrimination (D) as follows:

Item discrimination

$$D = \frac{H - L}{\frac{1}{2}N}$$

Where N = total number of learners from high and low score groups.

The highest value of D is 1 if all the learners in the H group got the items right and none in the L group got it right.

Normally, an item discrimination value of 0.3 or higher shows that the item

distinguishes between bright and weak learners.

Example: If 7 out of 10 in the upper group answered correctly and 3 out of 10 answered correctly in the lower group, then D is 7 minus 3, divided by 10 = 0.4.

The above is not that easy especially for a beginner. Worse still, even after much struggle to come up with a good test, you will always find that either the item was not that good or some bitter criticisms will come from those affected by test results (e.g. parents). This further calls for patience, an imaginative mind and lots of practice in item writing.

Conclusion

We have seen that a test writer requires to have:

- Knowledge of the structure and goals of the curriculum which, in the case of Science, would also include science process skills;
- Knowledge of the subject matter;
- Knowledge of the characteristics of the learner;
- Skills in test construction and correct interpretation of test scores;
- A creative mind;
- Patience to sit for long hours and also accept criticisms, etc.

Appendix I

Action Verbs to Use when Writing Behavioural Objectives

For knowledge	<ul style="list-style-type: none">• define• state• list• name	<ul style="list-style-type: none">• write• recall• recognize• label	<ul style="list-style-type: none">• underline• select• reproduce• measure
For comprehension	<ul style="list-style-type: none">• identify• justify• select• indicate	<ul style="list-style-type: none">• illustrate• represent• name• formulate	<ul style="list-style-type: none">• explain• judge• contrast• classify
For application	<ul style="list-style-type: none">• predict• select• assess• explain	<ul style="list-style-type: none">• choose• find• show• demonstrate	<ul style="list-style-type: none">• construct• compute• use• perform
For analysis	<ul style="list-style-type: none">• analyze• identify• conclude• differentiate	<ul style="list-style-type: none">• select• separate• compare• contrast	<ul style="list-style-type: none">• justify• resolve• break down• criticize
For synthesis	<ul style="list-style-type: none">• combine• restate• summarize• derive	<ul style="list-style-type: none">• argue• discuss• organize• conclude	<ul style="list-style-type: none">• select• relate• generalize
For evaluation	<ul style="list-style-type: none">• judge• evaluate• determine• recognize	<ul style="list-style-type: none">• support• defend• attack• criticize	<ul style="list-style-type: none">• avoid• select• choose• identify

The following are open to a variety of interpretations and therefore are **not recommended**:

to know	to appreciate	to enjoy
to understand	to fully appreciate	to believe
to really understand	to grasp the significance of	to have faith in

from: Kenya Institute of Education, Primary Teacher Education, Draft Teaching Guide for Science, 1985.

Appendix II

Science Process Skills and Attitudes

- | | |
|------------------|--|
| Skills | <ul style="list-style-type: none">• observation and recording• asking questions• sorting and classifying• collecting data• interpreting data• making predictions• drawing conclusions• reporting and exchanging information• making hypotheses• experimenting (controlling variables)• measuring• manipulating/handling materials |
| Attitudes | <ul style="list-style-type: none">• curiosity• self-confidence• cooperation• responsibility• interest• practical approach to solving problems |

from: Kenya Institute of Education, Primary Teacher Education, Draft Teaching Guide for Science, 1985.

Appendix III

Test Construction Guidelines

The following are standard guidelines applicable to the preparation of all assessment instruments.

1. All items of similar format should be grouped together.
2. Arrange items in order of difficulty.
3. Items should be well separated by leaving enough blank space between items for ease of reading.
4. Items should be completed on the same page where they begin.
5. Illustrative material should be placed above the items that refer to them.
6. Items should be randomised.
7. Select a format for answers.
8. Student data should be recorded.
9. Instructions should be reviewed.

10. Proof-read the test.

Guidelines for Constructing Essay Questions

1. Relate each question to an instructional objective.
2. Give adequate time for each item.
3. Indicate level of details required.
4. Use essay items only where objective items cannot be used to evaluate.
5. Allow no options so all candidates should do same items.
6. Evaluate the candidate in terms of his presentation rather than the position he takes in answering questions.
7. Define the tasks clearly and unambiguously.
8. Do not begin essay items with such words as »who«, »when«, »reason«, »evaluate«, etc. as these focus on recall. It is better to start with words like »compare«, »contrast«, »give examples«, »reason«, »evaluate«, etc.
9. You should know the processes you want the student to use in answering the question and select the appropriate verbs.

from: M.B. Ogunniyi, Educational Measurement and Evaluation (Hongkong, Wing Lee Printing Co. Ltd.), 1984.

3.5. Technical Efficiency in Constructing Test Items

Joseph M. Khakame, Kenya National Examinations Council

Introduction

Technical efficiency in construction of test items can be defined as a process of ensuring that a test item is reliable, valid and discriminates appropriately among ability levels of the examinees. This definition is based on the assumption that the item will form part of a test intended for Norm Reference Testing (N.R.T.). It is categorical that the test be for N.R.T. because the purpose for which the scores of such a test are used is basically selection and placement. In the Kenyan case, the Kenya Certificate of Primary Education (KCPE) examination is both terminal for the majority of examinees and for selection and placement of most of the examinees of higher ability.

Kenya has for many years since the early 1970s been striving to ensure that the examinations whose results are used for selection and placement adequately meet these requirements. In addition to these requirements, the examinations are used for certification. These examinations are:

- the Kenya Certificate of Primary Education (KCPE) (Certificate of Primary Education before 1985);
- the Kenya Certificate of Secondary Education (KCSE) (Kenya Certificate of Education before 1988);
- the Kenya Advanced Certificate of Education (KACE) which was phased out in 1989.

This paper will focus on the KCPE examination and the Science paper in particular.

Types of Tests

The tests to be discussed will be of achievement more than of aptitude. It is, however, difficult to distinguish between an achievement and an aptitude test. Aptitude tests are predominantly used to predict the future learning of a candidate. Whereas we categorise the KCPE tests as achievement, the purpose for which they are used make them have considerable resemblance to aptitude test items. Most of the items in these tests are ability-based and, for this reason, the tests have departed from the dichotomy of achievement versus aptitude testing. A follow-up study by the Examinations Council to check and confirm the predictive validity of these tests was carried out. The examinees who scored high on the KCPE Mathematics paper of 1985 performed extremely well on the KCSE Mathematics paper of 1989. This has proved that although the KCPE tests are predominantly achievement tests, they have high predictive validity.

There are many forms of tests which can be used to measure the achievement of examinees who have gone through a prescribed course of study. These are:

- essay-type (free response)
- objective type of which there are various types
 - multiple-choice
 - matching
 - true or false
 - structured (open-ended).

The essay-type items are used in the testing of English Composition and >Insha< (Kiswahili Composition). These two papers are scored by readers. This is a subjective scoring. The readers' knowledge of content and their moods can easily influence the reliability and validity of the scores. It is also argued that essay-type tests are better suited for higher ability testing. This is debatable but it is also clear that the testing is highly influenced by the candidates' verbal facility. The application of this mode of testing in KCPE therefore suits the purpose of the two papers except for the subjectivity of the readers.

The multiple-choice mode of testing is used in all the other papers for the KCPE examination. These include Mathematics, English Language, Lugha Ya Kiswahili, Home Science, Business Education, Geography, History, Civics, Religious Education, Art and Craft, Music, Science and Agriculture. There are three reasons for choosing this mode of testing at KCPE level. First, it is possible to sample a large part of the content to be tested. Second, its objectivity ensures that the scoring is accurate and in addition the responses of the candidates are least affected by verbal ability, other than the ability to read. Finally, scoring is possible by use of a machine which is faster than people and is not subject to bias or fatigue errors.

Although many people have a notion that multiple-choice type tests are used at elementary levels and the performance of the examinees is greatly influenced by guesswork, the truth of the matter is that this is a fallacy. It has been tested at university level, especially in the United States of America, and proved to be more effective than the essay-type examinations. Multiple-choice items can be used to test all ability levels. These are Knowledge (recall), Comprehension, Application, Analysis, Synthesis and Evaluation. This taxonomy of ability levels by Benjamin Bloom gives a basis for designing a test. It would be futile to design and write a test without a table of specifications in which the ability levels are clearly spelt out.

For each of the contents to be tested, specific weighting in terms of number of items per ability level must be determined. The formula used is not mathematical but takes into consideration the importance of the content to the learners, the amount of time the content is allocated in the curriculum, and the relationship of the particular content to the rest of the content. Below an example of a multiple-choice item strategy for high level testing is provided.

- Give a stimulus in the stem. This can be in the form of facts followed by a hypothesis.

- Four options are given where the examinee is to agree, disagree, show some doubt or doesn't know.
- For each option a set of ten to twenty facts are given where some of the facts are false.
- The examinee is asked to choose the option which best suits the hypothesis based on the appropriate accompanying facts.

The interesting thing about such a test item is that three of the options normally have supporting facts which will require a very high level of ability to be able to discount all the fallacies and finally choose the key from the three options. The fourth option is obvious for the examinees who do not know.

Test Item Analysis

Item analysis is a detailed statistical description of how a particular test item functioned when it was used in a test. The analysis provides information about the difficulty of the item, discrimination among abilities of examinees, the relative attractiveness of the options and deduced reasons for abnormal behaviour of the item.

Pre-test item analysis is a desirable evaluation of the suitability of an item. However, in Kenya the competitiveness of the system and the use of the end results of the tests prohibits its use.

The theory and practice of item pre-testing cannot be overemphasised. Kenya initially carried out pre-tests in the early 1970s, the results of which were used to set standards upon which the current KCPE examinations are based. The reasons why pre-testing is not used in this country are:

- the competitiveness of the education system makes pre-testing unreal;
- it is difficult to retrieve all the items after pre-testing;
- the results of a pre-test cannot be valid and reliable under the current set-up;
- the examinee cannot be motivated to take the pre-test to give valid results;
- there are many publications complete with answers and such pre-tested items can easily form part of the publications;
- the cost of pre-testing under the prevailing circumstances cannot be justified;
- the present method of test item analysis serves a very useful purpose. The results of this analysis, when used in the setting of a similar test, give the desired outcome;
- the impression of pre-testing to the general society would greatly injure the reputation of the Examinations Council.

The purpose for which such a pre-test could be used cannot be imagined and therefore lack of motivation on the part of the examinee would make the pre-test results invalid. The converse of this would be the impossible retrieval of the items from the examinees which would negate the original purpose and therefore the whole exercise would be futile.

Preliminary item analysis is usually performed on some candidates' scripts to check for the following before the actual scoring:

- a wrong key entered by the test writer,
- a wrong key entered in the computer system,

- a printing error that renders the key incorrect,
- an originally correct key made incorrect by the passage of some historical event.

Item Difficulty

The purpose for which a test is designed will dictate the nature of items to be used in the test. It is desirable to have a test where all the items are of low difficulty in an attainment test used to determine how well a group of examinees have learned a set of skills. If, on the other hand, the test is to be used for selection, as it is the case in Kenya, the items on the test must be of varying levels of difficulty. However, it is desirable in this case that a majority of the items should be in the middle difficulty range. The item difficulty is determined by the percentage of the examinees scoring the item correct. Take for example 24,136 examinees; when 12,068 of the examinees get an item correct, this will be 50% difficulty. From 50% to 60% is considered to be middle difficulty, 40% down to zero would be considered to be on the difficulty side. Again referring to the 1991 KCPE Science paper, there were 30 items, 16 of which were between middle difficulty and easy, six were slightly more difficult and eight were very difficult. The item difficulty results are used as a quick check to see which of the items behaved well. A more detailed analysis of performance is carried on from there.

Test analysis is carried out at the scoring stage of the KCPE examination. This analysis serves the following purpose: test for item bias.

It is very clear that Kenyan society has many social differences based on religion, culture and the geographical distribution of the examinees. There are many religious groups which are likely to influence the social set-up of various groups of people. At the test development stages, all items are carefully checked for bias. Any item which does not meet this stringent test is thrown out. However, there used to be one or two out of 100 in the 1970s that passed unnoticed by the experts.

This problem is no longer there. However, during this analysis, patterns of poor teaching emerge and this information is used in the writing of the backwash document normally referred to as the KCPE Newsletter.

Assessing Item Discrimination

Item discrimination is critical in a test if that test is to be used for ranking, selection and placement of examinees. This indicates the relationship between the high ability examinees' response to an item and the response of the low ability group in the same sample. Ideally, the sample should be divided into three groups and a comparison made between the top and bottom groups.

KCPE tests are of adequate length and the inclusion or otherwise of the item that is being tested for discrimination is not significant. The mean score for each of the two groups is least altered by the inclusion or exclusion of the score for the item in question. However, the Kenya National Examinations Council does not use this method of determining Item Difficulty Profile.

If we take a sample of 24,136 examinees, and divide it into three groups according to the ability ranges on the test, this will give us 8,045 candidates per group. It is assumed that the question to be analysed is of middle difficulty; therefore the middle group is expected to remain so. Let us take item No. 24 on the 1991 test in Science and Agriculture.

24. Sankuri classified some of the animals he had seen during a visit to a national park into two groups using a certain characteristic:

Group I	Group II
Lizard	Mouse
Frog	Ostrich
Crocodile	Leopard

Which one of the following shows the characteristic that Sankuri used?

Group I

- A. Carnivore
- B. Lay eggs
- C. Body temperature varies
- D. Do not care for their young.

Group II

- Herbivore
- Bear their young alive
- Body temperature constant
- Care for their young.

Sample statistics: N = 24,136 × 14,78 - SD = 4.48; Max. Score = 30

Table I

ITEM No. 24	RESPONSES				D-INDEX
	A	B	C*	D	
Candidates & % Choice	1,454 (6.01%)	4,064 (16.83%)	12,519 (51.86%)	5,975 (24.74%)	0.55
Mean Score (\bar{x})	10.39	12.17	16.46	14.22	

*Key

The KNEC method of determining the **discrimination index** is based on the dichotomy of correct and wrong responses made by examinees in the sample. Table I above shows the responses where the key represents correct response while the three distractors represent the wrong responses. Candidates who did not respond to the item (omits) have not been included in the table. To determine the index, the following formula is used:

$$D - \text{Index} = \frac{\text{Mean Score } (\bar{x}) \left(\begin{array}{l} \text{for examinees getting} \\ \text{correct response} \end{array} \right) - \text{Mean Score } (\bar{x}) \left(\begin{array}{l} \text{for examinees getting} \\ \text{the wrong response} \end{array} \right)}{\text{Standard Deviation(SD)}}$$

Using the results from Table I:

$$D - \text{Index} = \frac{16.46 - (\bar{x} \text{ of groups A, B, D, and omits})}{4.48} = 0.55$$

This is an item of middle difficulty, and the proportion of examinees scoring correct come from the high ability group (\bar{x} = 16.46) and those who were wrong are of average ability and below, that is \bar{x} = 14.22, 12.17 and 10.37 respectively.

Importance of Discrimination Index to an Item Writer

For example, if the mean score for the examinees getting the item correct was 12.00 and the mean for the candidates getting the item wrong was 16.40 from the formula:

$$\frac{12.00 - 16.40}{4.48} = 0.98$$

An item of this kind would indicate that there is something drastically wrong with the question in any of the following ways:

- the high ability candidates ignored the facts and chose to reason, but their reasoning did not have a sound scientific basis;
- the item has a bias which is likely to be fallacious and this obscures the facts which are likely to be very trivial;
- the item is so difficult that most candidates guess but in the process the higher ability candidates try to reason and this distracts them from the key.

An analysis of all items for item discrimination gives a guide to the setters to come up with similar items for the future tests that will:

- have the desired discrimination among candidates by ability levels. This is an essential requirement for a test whose results are intended for selection and placement;
- allow the low ability candidates to score on a few items. This is necessary because each examinee who takes the KCPE examination gets a certificate that shows attainment grades. In the Science and Agriculture paper 1991, there were four Science and eleven Agriculture items respectively that were easy. This gives a total of 15 questions which is 25% of the whole paper.

Critical Analysis of Test Item Discrimination Statistics

When item discrimination is done to a test, it is easy to put the print-out on the shelves and wait for the next year to carry out a similar analysis. However, if the findings of the analysis are critically looked at, very good use of the findings can be discovered.

The subject specialist and the item writers should do this analysis to be able to use the findings fully. Item No. 2 of the 1991 Science and Agriculture paper is analysed in detail.

2. As Kizito was adding kerozene to his lit stove, the whole stove caught fire. The MOST effective way to put out the fire would be

- A. covering it with soil
- B. beating it with a leafy branch
- C. covering it with a piece of cloth
- D. pouring water over it.

Table II

ITEM No. 2	RESPONSES				
	A*	B	C	D	D-INDEX
Candidates & % Choice	8,984 (37.21%)	1,556 (6.44%)	10,638 (44.06%)	2,874 (11.79%)	-0.02
Mean Score (x)	15.33	13.08	15.18	12.71	

* Key

This question was not discussed in the KCPE Newsletter because it was scored correct by more than 30% of the examinees which was the criterion for the choice of question appearing in the backwash document.

Item No. 2, however, is of interest for item writers because of the distribution of candidates choosing the key and those choosing a popular distractor.

First, the distractor C was chosen by 44.06% of examinees of above average ability, whereas the key A was chosen by 37.21% of examinees of above average ability. All other options were chosen by candidates who were of below average ability. It is apparently clear that this item is difficult. The two options A and C were the most plausible choices and since the item asked for the MOST effective way, both options can be effective depending on the circumstances. In some life situations, soil is hard to come by e.g. a candidate from an urban setting may not be able to gather 1 kg of soil in one hour by which time the fire would be completely out of control. If the same examinee under the urban conditions was quick to think, a wet cloth would have been a solution, but the cloth in question was not indicated to be either wet or dry. An examinee could make that decision that the cloth is wet, therefore it can be used to put out the fire quickly covering the fire so that the supply of oxygen is cut off.

The item writers and the subject specialists should imagine they were the examinees and what they could do in a real life situation.

Most of the items on the Science paper were beyond simple recall and in fact even those which may have appeared to be recall items contained information that required high level ability skills.

Relative Attractiveness of Choices

When an item writer drafts an item with four choices, it is hoped that those examinees who know the correct answer will choose the key and the remaining examinees will equally opt for the three distractors. But, unfortunately, this is to a lesser extent the case. The analysis of the 1991 KCPE Science paper revealed that:

- out of the 30 items, 10 had one very plausible distractor which made the other two distractors implausible;
- in cases where many examinees scored correct, for example items No. 11, 19 and 20, the remaining examinees were attracted by one distractor.

Consideration of the ability ranges of candidates opting for the strong distractor shows that the item was based on misconceptions and fallacies, that is if such an ability is below average. Table III shows a summary of the 10 items.

Table III Summary of 1991 Science Test Items (Selection)

ITEM No.	% Scoring Correct	% Choosing Main Distractor	Abilities
2	37.21	44.06	Low
6	10.90	52.80	Above average
8	37.18	34.02	Low
11	45.52	29.08	Mixed abilities
18	18.61	47.13	Mixed abilities
19	66.16	26.06	Low
20	68.71	23.84	Low
27	15.99	51.26	Mixed abilities
28	10.70	65.89	Mixed abilities
30	34.81	55.10	Low

The strong distractors should be carefully analysed so that similar distractors are not inadvertently used in future on similar items. Strong distractors where the majority of the examinees have low ability can be ignored.

All this data is available to the six setters of the paper. It is up to the setters to study these findings so that when they write items, they follow the set standards and only use the responses in the past paper as a guide. It is logical and reasonable to assume that a response distractor which attracted many examinees in the 1991 KCPE examination of above average ability must be either biased or the question was ambiguous. These types of responses are critically analysed and the conclusions are used as a guide in generating responses to the new items for a future examination.

In conclusion, the reason for avoiding to discuss or even mention reliability and validity statistics here is that these two features are derived from the mean, standard deviation and correlation relationships of a sample. They are the factors which are used in test item analysis whose results therefore fulfil the requirements of validity and reliability. The Examinations Council has over the years developed a system where item writers are trained before they can be used in the actual process of examination item writing. The current system of test item analysis is not fully developed.

There is still room for increasing the statistical features in the test item analysis, and the processes involved in test item analysis may be added to the training of the item writers in order to produce better test items in future.

4. APPENDICES

Int. J. Educational Development, Vol. 10.
No. 1. pp. 69-82. 1990
Printed in Great Britain

0738-0593/90 \$3.00+.00

Pergamon Press plc

4.1. Examinations Policies to Strengthen Primary Schooling in African Countries

THOMAS OWEN EISEMON

Centre for Cognitive and Ethnographic Studies, McGill University

Abstract - Examinations can be a powerful instrument for influencing the kind of instruction students receive, what they learn at school and how they will use school knowledge and skills in daily life. Attention is given to the examinations administered to African primary school students which are the basis for selection for secondary education. Findings are drawn from recent studies in two East African countries, Burundi and Kenya. These studies suggest: (1) that the scope of the school leaving examinations should be increased and should emphasize assessment of knowledge in the domains of science, modern agriculture, health science and nutrition; (2) greater use should be made of vernacular languages for testing to encourage explanatory instruction especially in scientific subjects; and (3) examination items should be constructed to measure high level cognitive skills of importance for performing practical tasks involving applications of modern scientific knowledge.

INTRODUCTION

This paper examines the use of school leaving examinations for strengthening the internal and external efficiency of primary education in African countries. It will focus on examinations policies in three key domains, those relating to: (1) the scope of examinations; (2) the language of testing; and (3) the construction of test items. It will suggest how manipulation of examinations policies may influence the kind of instruction students receive, and what they learn and may retain from schooling. Examinations can be a powerful and positive instrument of educational policy if their impact on instruction and learning is better understood. That is the purpose of the paper. Interest in using examinations for improving primary schooling in African countries has increased in recent years for many reasons; chief among them the recognition that examinations reform is a relatively low cost, effective strategy for stimulating educational change (Heyneman, 1986, 1987).

In contrast to primary schools in western countries with well developed systems of secondary education where the first years of schooling are considered crucial and are the object of much attention, schools in many parts of Africa tend to allocate scarce instructional resources to the upper stage of primary schooling which prepares students for highly selective secondary school entrance examinations. Better trained, more experienced teachers are often assigned to teach the final years. Students in the upper stage of primary schooling may receive more hours of instruction and are more likely to study from textbooks which are either distributed to them or which their parents purchase. Consequently, manipulation of examinations policies may not change the present allocation or perhaps more accurately, the misallocation of instructional resources. The intention of examination reform is, instead, to change the way these resources are used resulting in subtle changes in teaching and in the cognitive outcomes of primary schooling.

Educational researchers and policy makers in Africa and elsewhere have been ambivalent in their thinking about national examinations. Examinations are often viewed as the cause of poor quality primary schooling as well as the source of urban/rural, regional, gender and other disparities. This is contested by advocates of examination reform who have shown that well managed systems with the research capacity needed to design and assess test instruments

can not only produce examinations that avoid the usual biases but also provide feedback to schools to improve student achievement (Somerset, 1983, 1987; Savage, 1985). The Kenya National Examinations Council has been described as an exemplar of such a system (Somerset, 1983; Eshiwani, 1986).

THE USE OF EXAMINATIONS FOR EDUCATIONAL IMPROVEMENT

National examinations have a long history in developing countries with a colonial experience that is important to review. India was the first to experiment with examinations as a strategy for improving the quality of post-secondary education. These were not the national examinations we know today but external examinations set by universities established for the purpose of accreditation and examination which did not offer instruction. The three original Indian universities at Bombay, Calcutta and Madras founded in the late 1850s along the lines of the University of London had this function. Their purpose was to give uniformity to the first degree programs of affiliated colleges, increase standards and restrain the growth of higher education. By the 1880s, it was acknowledged that the affiliating/examining model had failed in all of these respects (Ashby, 1966). Lord Curzon's efforts at the beginning of this century to reform Indian higher education by strengthening the affiliating and examining functions of the universities that by then had become centres of sedition did not have greater success (Basu, 1974). Irregularities in affiliation and examination, the result of pressures from caste, communal and linguistic groups for expanded access to higher education, plague the Indian university system to the present day (Altbach, 1972; Eisemon, 1982). The lesson that Britain drew from this experience, and France from a somewhat similar one in Vietnam, was that educational expansion had to be controlled from the outset; controls could not be successfully imposed at a later time.

The lesson was applied in Africa where formal education for the indigenous population developed in a presumably more logical, more planned way, i.e. from the bottom up under the close scrutiny of colonial authorities. Use of external examinations set and marked in the metropolitan countries and administered to African students desiring higher education either abroad or at the collegiate institutions established in Africa after the Second World War effectively controlled access to further education and ensured success of initial efforts to implant metropolitan standards. Newly created institutions of higher education were also affiliated to metropolitan universities that, for a time, awarded degrees to students who undertook their undergraduate training in Africa. During the period of decolonization and the first years of independence, the external examinations required for university studies were partially localized. This was done by either adapting overseas examination papers for administration in Africa and increasing African participation in marking papers, or in the case of British West Africa, by establishing a regional examinations council to set and administer examinations that followed the metropolitan model of how external examinations should be conducted.

It is remarkable, in retrospect, how well these systems of external examination worked: so well that they were retained many years after independence. Overseas examinations for university entry are still taken by a large number of African students usually attending prestigious private secondary schools who hope to do further studies abroad. In many African countries, the overseas examinations continue to set the standards for measurement of achievement in academic subjects and for good examination practice, generally.

The creation of national systems of primary and secondary education and eventually higher education after independence led to the establishment of national systems of examination in many African countries or to mixtures of national and local examinations systems in the larger countries like Nigeria. There is great diversity in the ways these have been established. Perhaps the most typical model is an examination unit located in a national or state ministry of education, though many like the Kenya National Examinations Council have autonomous status. There is even greater diversity in the scope of the functions of these units. In Kenya, the Council is responsible for almost all examinations including those administered at teacher training colleges and most pre-university professional and technical examinations as well. That is atypical.

What is common is that few examinations units have the capacity to do more than set, administer and mark examinations, and announce results. They function for the purpose of allocating educational opportunities and in this respect, most are unlike their predecessors in the colonial period which had a broader mandate to strengthen the quality of instruction. The responsibilities of African examinations bodies are, however, much more complex. There are more levels and subjects of instruction to be examined, more students taking examinations, greater political pressures on the autonomy essential to the examination process and few resources needed to innovate on existing practices. The implication of this is, simply, that examination reforms at the primary or other levels of the educational system will be difficult and expensive to introduce, and harder to institutionalize than the examinations focused strategies of qualitative improvement characteristic of the colonial period. But the centrality of examinations to instruction and learning merits attention to examinations policies and to strengthening the capacity of examinations units as an element of any package of qualitative improvements which may also include better facilities, more textbooks, support for more frequent school inspection, in-service teacher training, and so on.

CURRICULAR INNOVATIONS AND THEIR IMPLICATIONS FOR TESTING

Major investments have been made by the World Bank and other donors and by African governments in developing capacities to prepare, evaluate, publish, distribute and to train primary school teachers in the use of instructional materials. Interest in textbooks has been accelerated by the findings of international studies of educational achievement which have stressed the contribution of textbooks and other qualitative inputs to student learning in developing countries (Heyneman and Loxley, 1983; Heyneman *et al.*, 1981). A great deal has been accomplished in this regard, more than is generally recognized. Twenty years ago, most of the textbooks used at the primary level were either imported from metropolitan countries or, in the case of vernacular language primers, published locally by religious groups. Today, many African countries have textbook publishing industries of varying scale, and a cadre of trained textbook writers and illustrators. Larger countries like Kenya and Nigeria have significant private sector involvement in textbook development and publishing, though ownership, management and technical support is more often foreign than indigenous (Eisemon *et al.*, 1986).

This capacity has enabled African countries to make substantial efforts to adapt and to expand the content of primary schooling. What that implies differs from country to country. There are, nevertheless, three general trends. The first relates to the increasing importance that has been given to the teaching of vernacular languages. Here, again, there is much variation. In Tanzania and Rwanda, vernacular languages are used for instruction throughout the primary cycle in most rural schools. In Burundi, the local language is used in the lower stage of the primary cycle and a transition is made to the metropolitan language in the final years. In Kenya, local languages and/or Kiswahili or English may be used in the lower stage while English has been retained for the upper stage. In most countries, the teaching of vernacular languages has been assigned a central place in the primary school curricula and in the rhetoric of educational reform. However, metropolitan languages continue to be used for upper primary and secondary schooling in many African countries.

The second concerns the teaching of practical subjects such as health, nutrition, home economics, agriculture and pre-vocational skills. These subjects were also emphasized in primary schools in the colonial period (see, for example, Bude, 1985; Foster, 1965). Their revival as a strategy for fostering rural development has been prompted in part by recognition that the possibilities for expansion of secondary education and formal employment are limited. Most students who attend primary school will return to their families and whatever knowledge and skills they have acquired will be put to use in a rural environment, either in agriculture or in some form of employment they will create for themselves. The primary school curricula has been enlarged in consequence. In Kenya, for instance, twelve subjects are taught in the final year of primary school!

The third tendency has been to encourage what was described in the 1960s as discovery or inquiry oriented instruction, or what is still known in some parts of East Africa as the 'new approach' to primary school teaching (Eisemon *et al.*, 1986). In the late 1960s, the United

States Agency for International Development (USAID), UNESCO, the Nuffield Foundation, the British Council and other donors initiated curriculum development programs in Africa which were mainly aimed at the improvement of primary and secondary school teaching of science and mathematics (Yoyole, 1985). These employed, what were then, the latest theories of instructional innovation in Western countries emphasizing development of high level cognitive skills.

Each of these trends poses a major challenge to established examination policies and practices; which subjects should be examined to ensure coverage of the mandated curricula, how should practical knowledge and skills be measured, in what language should testing be carried out, and how can examinations be developed to assess students' conceptual understanding of subject-matter as opposed to mere recall of disparate facts on which they have been drilled?

Testing to increase coverage of the school curricula

Achievement tests have the dual purpose of measuring what students have learned to ascertain whether certain minimal levels of performance are obtained, and selecting some students for further education. These objectives are not necessarily incompatible. But often if there are few opportunities for further education to be allocated, selection guides the construction of examination papers and test items. If domains of content are left off such tests, they are likely to be neglected in instruction. A test with good predictive validity, but poor curriculum or instructional validity may select good students for secondary school, but undermine the educational objectives which are intended to benefit students who are not continuing their formal education beyond primary school. The concepts of curriculum validity and instructional validity put the burden of proof on the examining authority to show that there is a reasonable relationship between examination content, the intended curriculum, and what is actually taught.

We have studied the impact of examination design on curriculum coverage at the primary level in Burundi and Kenya (Eisemon, 1988; Eisemon *et al.*, 1989a, b). Both countries have highly selective systems of secondary education, entrance to which is dependent on student performance in a rigorous examination administered at the end of the primary cycle. In Kenya, about a third of the students taking the Kenya Certificate of Primary Education Examination go on to some form of secondary education. In Burundi, less than 10% do so. Compared to Burundi, primary school instruction in Kenya is poorly monitored to ensure implementation of the curricula. Many rural schools are seldom inspected. Moreover, there are wide variations in the factors that influence coverage of the school curricula like trained teachers, textbooks, classrooms and instructional space for teaching many subjects, but especially compulsory practical subjects. In Burundi, there is much less variability among primary schools and instruction is closely supervised. For instance, school directors are required to make one hundred and fifty classroom visits during the school year and file reports with school inspectors.

Students taking the *concours national* in Burundi are examined in only four academic subjects and student marks are weighted in favour of tests of French reading comprehension and grammar. The examination is designed mainly from the perspective of selection for secondary education. Not surprisingly, we have found that the instruction given in the upper stage of the primary cycle often deviates from the curricula despite, or perhaps because of, close teacher supervision. Teachers frequently skip lessons in compulsory subjects like agriculture and home science that are not examined (Eisemon *et al.*, 1989a).

Kenya's Certificate of Primary Education examination has been designed to measure student achievement in all subjects in the school curricula, and the various papers are weighted equally in calculating final marks. The introduction of the 8 + 4 + 4 system in 1985 lengthened the primary cycle by an additional year. Its most important feature, though, was the emphasis on Kiswahili and practical subjects. Although the government produced a revised syllabus and new textbooks, very little was new in these curricular reforms except the emphasis on pre-vocational training. The major innovation was that what was supposed to have been taught for many years was to be examined. The Kenya National Examinations Council was given the

task of setting new papers for health (previously taught and examined as integrated science which is to say that little importance was given to it), Kiswahili (not examined, and therefore, often not taught in the upper stage of primary schooling), agriculture (hitherto a small part of the science syllabus), and other subjects like art and crafts (on the basis of project work). These subjects were made compulsory papers. The changes were widely criticized in the public press and resisted by many educators (Eisemon, 1988). Implementation created many practical problems; facilities had to be constructed, texts produced, classrooms equipped and teachers trained for the new syllabus. Nevertheless, the effect was dramatic. Syllabus coverage, our studies indicate, was increased in spite of a great many difficulties in both urban and rural schools (Eisemon, *et al.*, 1986; Eisemon, 1988).

A strong and consistent, if not particularly insightful, finding of most research on effective schooling is that differences in the scope and amount of instruction received produce large variations in student performance. Typically, reducing such variation and raising achievement levels is seen as a school management problem to be remedied by closer teacher supervision, more frequent school inspection and by measures that will ensure a more equitable distribution of instructional resources like textbooks. The impact of examinations policies on students' opportunities to learn what primary schools are supposed to teach is less well recognized.

Coverage of the school curricula is a serious problem in primary schools in many African countries. While the scope of examinations can be increased to foster greater coverage, there is often not enough time or instructional resources to cover the subjects that ministries of education have prescribed, or to teach them in the innovative ways that curriculum developers expect (Eisemon *et al.*, 1989a). Ministries have various policy devices to deal with the problem; for example, revision of the scope of the school curriculum to bring intentions closer to the realities of implementation. Unrealistic expectations are in part responsible for low levels of test performance. In Burundi, for example, average scores for recent mathematics and French comprehension and grammar papers in the *concours national* are only slightly above the chance level. A closer relationship between examining authorities and those responsible for curriculum development and teacher supervision is required.

Testing practical knowledge and skills

Some of the most influential instruction that takes place in primary schools in terms of the ways school leavers think about and perform tasks in daily lives, is given in subjects that are often not examined or are poorly covered in secondary school entrance examinations; for instance, health, nutrition, home economics, principles of modern agriculture and science, especially biology. Testing in academic subjects like language arts and mathematics is also seldom designed to assess skills and knowledge of practical importance.

Information processing skills are important to a wide range of practical activities; for example, comprehending instructions on how to do some practical task such as administering medicines or applying fertilizers. The ability to encode, decode, recall and make text based inferences is not sufficient for comprehension. Readers must make inferences from their prior knowledge and experience in understanding these texts. Comprehension of such texts is one of the mechanisms through which schooling facilitates adoption of modern health and agricultural practices based on interventions derived from modern science.

Processing procedural information requires performing complex cognitive tasks which make literacy functional. Procedural texts have many features that do not enhance comprehension. Typically, they are brief texts with a high density of information expressed in many conditional statements (Eisemon *et al.*, 1987; Eisemon, 1989b). Crucial information is seldom marked with text cues aside from those setting apart different kinds of information. The structure of this genre of printed information is highly varied. There are relatively few conventions to which discourse must conform that might facilitate information processing. Although their purpose is instructional, these texts often assume familiarity with the subject and some related expertise. In many African countries, procedural texts require comprehension of information presented in two or more languages for correct inferences (Eisemon and Nyamete, 1990).

The mathematical problem solving tasks imbedded in procedural texts such as calculating the amount of a product that must be purchased or how much of it is to be administered are difficult to perform with the information provided. Units of measure are frequently unfamiliar and necessitate conversion. For this reason, adults rarely perform calculations in everyday life compared to how often they estimate solutions to practical problems. Abilities in estimation, of course, have a relationship to skills in calculating correct answers. But these skills are not necessarily a proxy for each other (Eisemon, 1989b; Eisemon *et al.*, 1989a).

Schooling affords students little exposure to text genres and numeracy tasks they will be exposed to as adults and, in consequence, little opportunity to develop the necessary skills (Eisemon, 1989b). Moreover, conventions of testing poorly simulate naturalistic tasks. Reading comprehension tests usually present students with narrative or descriptive texts and a set of tasks that involve recalling or inferring information from the texts as if comprehension does not involve using prior knowledge. Most comprehension tasks do not require integrating information presented in more than one stimulus text and testing of literacy and numeracy skills is done independently. Mathematics instruction and methods of examining achievement emphasize correct answers, penalize students for good 'guessing' through marking practices and powerful distractors, and make extensive use of narrative texts for constructing problem solving tasks ('story problems').

Elsewhere (Eisemon, 1988, 1989b; Eisemon *et al.*, 1989a, b) we have described alternative and more naturalistic approaches to assessing literacy and numeracy skills in primary school students. Stimulus texts were designed to incorporate many features of procedural texts and comprehension and numeracy tasks constructed to elicit prior knowledge acquired from instruction in science and other subjects. Similar texts and tasks were used to measure adult literacy and relate it to important outcomes of schooling. Among Burundian and Kenyan farmers, for example, functional literacy scores had a larger effect on farm output than years of schooling and most other indicators of cultivators' skills and production characteristics, including expenditures on fertilizers and other modern production technologies (Eisemon, 1989a). While there is much scope for testing practical skills in academic domains, many complex issues are raised. These have to do with techniques of measurement, assumptions about how functional knowledge and skills should be assessed and, ultimately, about the way language arts and mathematics are taught.

In many African countries, there are strong pressures to vocationalize basic education, resulting in programs to prepare students for self-employment. In contrast to technical training which has as its primary objective the development of productive skills leading to levels of measurable expertise, the aims of pre-vocational instruction are more modest and more ambiguous. Most pre-vocational instruction has two foci; the development of skills that may be useful in a wide range of production situations as well as those which are associated with particular kinds of work whose mastery is in some sense a prerequisite for subsequent technical education, formal or informal. Because such instruction occurs in the context of basic education and employs methods drawn from the teaching of academic subjects, pre-vocational education has the additional purpose of providing experiences to facilitate the transfer of cognitive skills from academic to practical domains.

Measuring outcomes of pre-vocational education is complicated by many factors, not least is the difficulty in deciding what production knowledge and skills should be tested, and how this should be done. In academic subjects, knowledge and skills are hierarchical and cumulative, at least from the standpoint of curriculum and measurement specialists. And there are widely accepted methods for assessing task performance. However, there is little to guide testing of outcomes of pre-vocational education.

Improving testing of outcomes of pre-vocational education requires understanding: (1) the knowledge and skills important to productive activities and their relationship to schooling; (2) key production tasks associated with these activities; (3) what constitutes competent task performance; and (4) how expertise is acquired. In brief, it necessary to look outside school to develop strategies for assessing the effectiveness of practice studies.

In Burundi and Kenya, agriculture and crafts are the principal subjects of pre-vocational

education. We have studied how primary school children are taught these subjects as well as how production knowledge and skills are used in everyday life (Eisemon *et al.*, 1989b; Eisemon, 1989b). School instruction in agriculture and crafts share certain characteristics. Typically, production knowledge and skills are imparted by individuals without training, experience or much interest in the domain of instruction. Thus, students have few opportunities to observe expertise. Practical work is emphasized but little importance is given to competent task performance. Progression from one topic of instruction to another is not based on demonstrations of task mastery. While topics may be ordered in the syllabi according to some taxonomy of complexity, there is often a great deal of redundancy indicating that the skills may not be hierarchical.

Outside school, children acquire agricultural and craft skills in very different ways. Children observe adults engaged in productive activities from a very early age and through this, encode important information about production processes. They acquire production skills through experimentation usually imitating adults. Learning of production tasks is spontaneous and often undirected. Children, having become proficient in one task, learn another gradually expanding their repertoire of skills. Learning agricultural and craft skills always involves working with objects of value to produce things of value.

In so far as craft skills are concerned, our studies suggest useful strategies for measuring skills but also indicate that those implicated in expert performance are unlikely to be learned in school. The Kenya Certificate of Primary Education examination presently tests outcomes of craft instruction by requiring students to produce craft objects for assessment. A better method, that is more consistent with the aims of pre-vocational education, would be to test student performance of tasks fundamental to expertise (Eisemon *et al.*, 1988). Efficiency, for example, is fundamental to expertise in many craft occupations like soapstone carving, an important form of self-employment in western Kenya. Assessment tasks can be constructed to measure students' efficiency in selecting stones for carving or in the choice of tools and subject matter. But such expertise is unlikely to be developed through formal instruction. Soapstone carving is probably like many other craft occupations such as carpentry and tailoring in these respects. Assessment of student products may provide few insights into the skills students have acquired that are involved in expertise or can be generalized to other practical domains. Improving present assessment practices engages the issue of whether these skills are best taught in school.

While many primary school students who are being taught crafts are novices, they begin instruction in agriculture with much production knowledge and skills. In fact, agriculture is not a pre-vocational subject for most students in rural areas. It is their present and future vocation.

Literacy and numeracy are one mechanism through which schooling seems to influence agricultural productivity. These skills affect capacities to effectively use modern agricultural inputs. Also involved is a knowledge of the scientific principles which underlie modern agricultural practices. Schooling affects the way farmers think about and practice agriculture (Eisemon, 1989a). Testing outcomes of agricultural instruction should emphasize how well students are able to apply the scientific knowledge they have acquired to practical production problems.

While variability is characteristic of most production situations, there are many production problems that exist across a wide range of situations - low soil fertility, poor drainage, crop infestation, and so on - and practices that are effective in dealing with them derived from modern as well as indigenous scientific knowledge. Farming expertise involves an ability to identify these problems, deduce their causes and select appropriate intervention strategies. Schooling should impart 'theories' and information that guide practical problem solving, especially involving the use of modern agricultural technologies. The effective use of hybrid varieties to increase yields is illustrative. This requires some knowledge of genetics and soil chemistry, however rudimentary.

The scientific principles of modern agriculture may be taught, or students expected to integrate what they have learned in science and agricultural classes (Eisemon, 1989a). In

Kenya, agricultural teaching at the primary level incorporates biology and other sciences and until recently, students were examined on the knowledge of principles of agriculture rather than their practical skills in, say, crop care. In Burundi, agriculture is divorced from instruction in science and it is not a paper in the national examination administered at the end of primary school. In order to investigate the effectiveness of agricultural instruction in Burundi, we have developed tests similar to the ones formerly used by the Kenya National Examinations Council to measure achievement in agriculture which assess problem solving skills using modern scientific knowledge (Eisemon *et al.*, 1989b). However, they are different in two respects. First, we have attempted to measure students' abilities to solve some practical production problems with indigenous knowledge and practices. Agricultural education in Burundi and most African countries condemns indigenous farming knowledge and practices by neglect, though such knowledge and practices like intercropping, particular patterns of crop rotation, crop fumigation and use of indigenous varieties are demonstrably effective in many production situations and some times superior to modern production technologies. Second, because agricultural education in Burundi, Kenya and other countries is offered in a metropolitan language, as is most science education, we have tested students in their mother tongue to better assess what they have learned in both subjects, and whether they have integrated this knowledge into their thinking. In sum, testing outcomes of pre-vocational instruction necessitates careful specification of what kind of impact schooling might have on domains of performance, taking into account the production knowledge and skills which children possess, and designing measurement strategies that capture crucial features of expertise.

Testing in the mother tongue to improve measurement of achievement

Language policies in many African countries where basic education is the responsibility of the national government, prescribe the use of a metropolitan language for instruction in the upper stage of the primary cycle and for examination in most academic subjects. This is seen as having several practical advantages, one of which is that in linguistically heterogeneous situations, learning and being examined in a metropolitan language is perceived to favour no particular linguistic group. In linguistically homogeneous situations like Burundi where French is the medium of most instruction in the final years of primary schooling and is used for most papers in the *concours national*, other justifications are offered that, essentially, favour the preparation of an elite for higher education *vis-à-vis* the competing needs of those who are unlikely to use French in later life. Examination policies, thus, determine language of instruction policies at the primary level just as much as they reflect those adopted for secondary and higher education.

Poor proficiency in the language of instruction is an impediment to acquiring knowledge and skills and, thus, depresses levels of performance, though not necessarily to the same extent. In other words, students with little grasp of the language of assessment may 'know' more than their test results reveal. There is much evidence of that in African studies comparing the performance of students in the mother tongue and in a metropolitan language (Zepp, 1982; Eisemon *et al.*, 1989b). Nevertheless, what performance differences reveal about how testing in a metropolitan language affects the measurement of student ability is not well understood.

In Burundi, we compared the performance of students in French which is used for the final two years of primary schooling and for the *concours national* as well as in Kirundi, the mother tongue of most students in that country (Eisemon *et al.*, 1989b). Tests were developed measuring reading comprehension skills and problem solving skills in the domains of science, agriculture and mathematics. As expected, students who received the tests in Kirundi scored higher on most tests than those who took them in French, although performance levels were low for all tests. The performance differences were largest for the tests of achievement in science and agriculture. In mathematics, however, students scored slightly higher in French.

Responses to question items were examined using monotone regression splines analysis to determine how language of assessment affects the performance of students of different ability (Ramsay and Abrahamowicz, 1989). In this analysis, ability was estimated from the student's response pattern for each test, and the estimate compared to the student's response to individual questions. In the reading comprehension and science and agriculture tests, the

performance of the most able students was most affected by being tested in French. Conversely, testing in the mother tongue did not increase the performance of students with less ability. These findings have at least two implications. First, students with poor proficiency in the metropolitan language learn so little that their achievement levels are almost unmeasurable. Greater attention needs to be given to raising the proficiency of these students in the metropolitan language. Second, mother tongue testing better estimates the knowledge and skills for those students who, presumably, are the most proficient in the second language. It may be useful for better identifying the most talented students for further education.

In mathematics, the French and Kirundi tests performed similarly in discriminating students. Since most of the test items were story problems, the similarity can not be explained simply in terms of the fact that the items required little language comprehension for correct solution. Most required complex representations of information for problem solving. The similarity in test results may be attributable to the use of the mother tongue for instruction in mathematics in the lower stage of the primary cycle which facilitates learning, especially for students with poor proficiency in French. In learning science in French or in understanding French scientific texts of the kind that were used to measure reading comprehension, students' difficulties may be compounded by lack of mother tongue scientific vocabulary and concepts. The lack of transfer of knowledge and skills from the metropolitan language to the mother tongue of African students is often noted (e.g. Berry, 1985). The importance of the development of the mother tongue for learning science in the metropolitan language is not given much attention in this connection.

These findings lead us to question the use of metropolitan languages for instruction and assessment in subjects like science, health and agriculture which have great practical importance for the majority of primary school leavers who do not continue their studies. There are several reasons why serious consideration should be given to expanding the scope of vernacular language assessment beyond the usual language arts papers, even if this is not accompanied by radical changes in the use of metropolitan languages for instruction. For one thing, it may provide a more accurate representation of the knowledge and skills primary school students have acquired and a better indication of the quality of education they have received. In addition, use of vernacular language questions to measure achievement in science, agriculture, health and nutrition may encourage more explanatory instruction. That, at least, might motivate many teachers to ensure that what is taught in a metropolitan language is really understood. It might also facilitate development of the student's mother tongue as a language of scientific discourse and, in doing so, increase capacities to use science in daily life. Indigenous languages will not develop as languages of scientific discourse unless they are used for instruction in science and practical subjects.

Testing to foster better teaching

Cramming is usually deplored in discussions of the 'backwash' effects of achievement testing. Still, it is necessary to distinguish between good and bad examination cramming. Bad cramming has three characteristics: (1) it involves drilling and accords little importance to learning activities requiring self-study skills; (2) it is focused on increasing student's exposure to possible examination items and correct answers and not on developing basic knowledge and skills, and (3) it results in distortion of the program of instruction in order to make time for extensive review.

Bad cramming may be an important though unintended outcome of the way examinations are constructed and administered. For instance, test items constructed to measure recall of textual information promote teacher reliance on choral recitation as an instructional strategy. Similarly, a highly randomized process of selecting examination items, though it may encourage teachers to cover the subject syllabi, also encourages unfocused examination preparation. Conversely, highly predictable examination items provide incentives for teachers to skip some lesson topics in order to increase time for review of other more important topics. Avoiding these extremes requires an understanding of teacher examination preparation strategies; i.e. how they decide what is likely to be examined, their beliefs about what guides the construction and selection of examination items and the methods they employ to coach students.

In Kenya, a great deal of information is made available to teachers about the content of the school leaving examinations. The Kenya National Examinations Council publishes sample papers, newsletters to teachers and, of course, many teachers are selected to mark the examination scripts under the Council's strict supervision. There is, in addition, an important indigenous cottage industry in the production of examination guides containing samples of questions used in previous examinations. The guides are purchased by parents and by teachers. Teachers, we have found, make extensive use of such guides in preparing students for the KCPE examination largely because they are suitable for class drills and written exercises (Eisemon *et al.*, 1988). The expanding sales and number of new editions of these guides provides some indication of the confidence that parents, students and teachers place in them.

Teachers select the questions that are most likely to be asked and drill students on the correct answers (Eisemon, 1988). Teachers 'work backward' from their knowledge of examination questions in constructing regular lessons, highlighting topics that are the most often examined and, again, drilling students on the correct answers. Teachers often do not vary either the form or the content of previous examination questions in their drills. They are presented *verbatim*. That may be the result of a poor understanding of how the examination items are constructed, i.e. of the knowledge and skills they are supposed to measure. Students do not, it seems, need to know *how* to answer the questions but rather *what* to answer. This may also reflect the need to cover as many examination items as possible. The students' exercise books into which these items are transcribed have the characteristics of examination guides; they are simply lists of questions and answers.

Examination preparation does not involve explanation of principles and concepts which students can use to organize knowledge that test items may be intended to elicit. It is not that teachers assume that important principles and concepts are understood from previous instruction. There is very little explanation given by teachers under other circumstances. Classroom discourse in academic subjects is characterized by vocabulary building, by introduction of new and unfamiliar terms in a language that is neither the student's nor the teacher's mother tongue. For this reason, teachers favour listing and fill in the missing word exercises in regular classes, and drilling the correct answers when coaching students. Such strategies are far from optimal in terms of student understanding. They persist since they are rooted in a complex of factors such as insufficient teacher understanding of subject matter, insufficient assistance from principals and others (inspectors, Ministry of Education staff development personnel), lack of textbooks and other instructional materials, lack of instructional time, pressures for examination success as well as the content and construction of examinations.

In Burundi we have investigated how teachers prepare primary school students for examinations, and the impact of teacher strategies and skills on student performance (Eisemon *et al.*, 1989b). Examination preparation practices involve; (1) increasing the amount of instruction students receive; (2) review of essential content; (3) frequent testing to identify performance deficits and build test taking skills; and (4) diagnosis of sources of student errors and selection of appropriate instructional methods to remediate them.

Teachers used various strategies to increase instruction prior to the administration of the *concours national* examination; lengthening the school day, substituting lessons and combining morning and afternoon shifts. Only a few teachers however, increased the length of the school day in the period prior to the examination. In most cases, this could not be easily done because many teachers and students resided at considerable distance from the schools. In some schools, the school week was lengthened by adding Saturday afternoon classes. As noted above, many teachers used class periods allocated for teaching pre-vocational subjects to teach academic subjects. These practices did not have much direct or indirect impact on student performance because the amount of additional instruction provided was insignificant compared to the effects of combining shifts. The number of times classes were combined was strongly related to student performance particularly for subjects like mathematics.

The impact of review strategies on student performance was difficult to assess. Creating more time for review of the syllabi may or may not be an effective examination preparation strategy, depending on the coverage of examination items, the predictability of item selection, the way the subject syllabi are constructed, and other factors. Teachers with a history of success in preparing their students for the *concours national* examination varied greatly in the amount of time they spent in lesson review. Some strictly adhered to the instructional timetable which allows only about three weeks for review. Others accelerated coverage of lesson topics in order to begin reviewing before the start of the third trimester when the examination is given.

What teachers review and how they review is, of course, more important than when they begin reviewing for examinations. The point is that successful teachers were not effective merely because they did more cramming. Nor, surprisingly, was student performance correlated with taking frequent mock examinations. In Burundi, these are set and administered at the cantonal rather than at the school directorship, school or classroom level. Moreover, the examining authorities do not produce sample papers or newsletters to guide teachers in preparing students as the authorities do in Kenya. Nevertheless, teachers do design examination-like learning tasks especially for review lessons. In some lessons we observed, these were simply correct answer drills. The teachers assumed that student examination success was dependent on their ability to recall correct answers to questions that might be formulated in many ways. An analysis of recent *concours national* examination papers suggested that this assumption was not entirely unfounded.

The most effective teachers, though, were those with the most insight into examination tasks and student performance. We presented teachers with sample examination questions and student responses. We asked them what students needed to know to correctly answer the questions, why they made certain errors and what could be done to better prepare the students. Teachers' responses were rated. Teacher ratings exhibited a high degree of variability both within the group studied as well as across the subjects of examination items. A few teachers scored highly on all measures, and a few very poorly. Teacher ratings for specific subject domains were correlated with student test scores and *concours national* results for science, mathematics and reading comprehension. The effect of teacher skills was stronger when opportunity to learn and school management factors were taken into account. What these findings suggest is that effective teaching is enhanced by teacher understanding of examinations that might be developed through in-service training. Another implication is that examining authorities should provide more guidance to teachers as to how they should prepare students for examinations.

Improving examination questions

Examinations should be viewed as determinants of the teaching students receive; as inputs, not just terminal outcomes of schooling. They influence how teachers select and organize knowledge for instruction. They also affect the learning tasks teachers construct from which students acquire and practice skills.

The expansion of primary schooling has favoured increasing reliance on the multiple choice format for examinations in many African countries. Multiple choice tests have the advantage of being machine scoreable and, thus, tests constructed in this format are easier and less costly to administer, and less subject to allegations of evaluation bias. Kenya has probably had more experience with such testing than most African countries, though a large proportion of the examination papers administered by the National Examinations Council are set in the more conventional essay mode and marked by trained evaluators.

In the early 1970s, major efforts were made to increase the proportion of Certificate of Primary Education examination items requiring 'reasoning' as opposed to use of descriptive information in answering questions (Somerset, 1983, 1987; Savage, 1985). Comparison of test items in the 1973 and 1976 examinations indicated that while the proportion of descriptive items declined from 74% to 23%, those involving some reasoning increased to 28% the remainder measuring lower level cognitive skills. The 1973 examination apparently did not require any reasoning for successful performance (Heyneman, 1986, p. 43).

Our interest in examination construction was prompted by work which examined how adults used literacy skills and scientific knowledge acquired in primary schools (Eisemon *et al.*, 1987; Eisemon and Nyamete, 1990). We found that adults experienced a great deal of difficulty in integrating knowledge acquired in school and from social experience in performing practical tasks involving administration of modern medicines and use of agricultural chemicals. In studies of Kenyan mothers' comprehension of texts for using commercial oral rehydration salts solutions, for example, well-schooled mothers with complete primary education who used these products often combined the treatment with administration of traditional purgatives. They did not understand how oral rehydration therapy worked and, thus, also relied on traditional treatments they could explain but which might worsen their child's condition.

Oral rehydration therapy is mentioned in the health syllabus and questions about it appear in the Kenya Certificate of Primary Education examination which are expressed in the multiple choice format; for instance: Your baby brother is suffering from diarrhoea and vomiting. Which one of the following would you give him to improve his condition? (A) plenty of fruit juice; (B) breast milk regularly; (C) plenty of water with a little sugar and salt; or (D) plenty of milk and porridge (KNEC, 1984, p. 119). (C) is the Correct answer although the administration of any fluids and nourishment is essential to recovery from diarrhoeal illnesses. When teachers we have observed prepare Standard VIII students for the examination, they usually have them recite the correct answer and transcribe the question into their exercise books for later study. The formula for oral rehydration therapy is to be committed to memory so that it can be recalled.

An experiment was conducted in a Nairobi primary school in which we asked teachers to prepare Standard VIII students for two sets of mock examination questions; one taken from the Health Science and Science sample papers prepared by the Kenya National Examinations Council, and the other covering similar topics that were revised to elicit high level cognitive skills. The revised questions were expressed in the multiple choice format of the sample papers (Eisemon *et al.*, 1988). The difference had to do with how the questions were constructed and what they required of students. For example, a question on knowledge of incubation, vaccination, immunization and modes of transmission of infectious diseases that in the sample paper merely required students to recognize the definition of immunization was changed so that students had to infer the correct answer. The premises were imbedded in a scenario that like the question concerning oral rehydration therapy, represented events in everyday life. But the 'wrong' answers were not plausible if a student actually understood that immunization usually confers immunity even when a child is exposed to contagious diseases.

The revised questions promoted significant changes in the ways teachers prepared students to answer them. Teachers spent more time on lesson preparation and read the relevant sections of the teacher guides and student textbooks. Their lesson plans emphasized explaining the meaning of new concepts and terminology. In the review lessons, teachers explained disease processes and various precautionary or treatment modalities. The teachers commented on how difficult it was to prepare for and teach the lessons in comparison to the review lessons that are normally taught. The students were then tested to ascertain what they had learned. Preparation for the revised questions enhanced performance on both tests. Think aloud protocols were obtained from a sample of students to examine how they answered the questions. The results indicated that the revised questions elicited explanatory knowledge of oral rehydration therapy and other health subjects, and that the students acquired this from the instruction they had received. Examination preparation had not imparted mere facts.

Improving achievement testing should not be limited to better examination administration, marking and reporting of results, though that is certainly important. The construction of examination questions needs to be improved as well. The psychometric properties of examination items do not necessarily reveal a great deal about how good a question is in eliciting meaningful demonstrations of competence or encouraging teaching that facilitates development of high level cognitive skills. These objectives should guide the construction of test items to which teaching efforts are addressed. Major changes in teaching practices are likely to result which, in turn, influence what is learned **and how** school knowledge and skills are used.

CONCLUSIONS

Primary school leaving examinations in many African countries, those which select students for secondary education, either cover only the academic subjects in the school syllabus or emphasize these subjects in weighting student performance. From the standpoint of the majority of primary school leavers who do not go on to secondary school and who will remain agriculturalists or combine agriculture with informal sector employment, the examinations should be designed in such a way that the subjects which should be the most important to them are taught and achievement measured. This means testing subjects like modern agriculture, health and nutrition and giving the marks obtained importance in examination results as is the case in Kenya. Different approaches to teaching and assessing learning in language arts and mathematics to foster acquisition of functional literacy and numeracy skills are also required.

The language of assessment influences the cognitive skills examination items measure and, in turn, affects teaching and how school leavers use information they have learned in school. Most assessment is carried out in a metropolitan language. However, primary schooling is the terminal stage of schooling for most children who will have little opportunity to develop or even retain proficiency in metropolitan languages after they leave school. Use of a vernacular language for assessment of student performance in scientific and practical subjects is suggested.

Finally, careful attention must be given to the ways achievement is measured. Teachers teach students to answer examination questions. If examination questions do not measure high level cognitive skills, teachers will place importance on students' recall of correct information for examinations. Changing test items to elicit high level cognitive skills associated with reasoning and problem solving and knowledge that is relevant to performing tasks in daily life may improve teaching and strengthen the effects of schooling as well.

A prerequisite for any successful strategy for improving examinations is a capacity to undertake research involving not only conventional psychometric studies of test characteristics but also investigation of how students perform examination tasks and how the examinations relate to the instruction they receive. Few examining authorities in rich countries have such capacity. But the need to acquire it may be greater for many African countries in which external examinations have a more important role in allocating educational opportunities and resources for improving the quality of education are more limited.

Acknowledgements - A draft of this paper was presented at a conference organized by the World Bank on Uses of Examinations and Standardized Testing in Africa. Lusaka. Zambia, 28 November-1 December 1988. Reference is also made to papers presented at another World Bank seminar on Using Examinations and Testing to Improve Educational Quality held in Kathmandu. Nepal. 1-3 November 1989. Some of the research reported here was supported by grants from the International Development Research Centre for studies carried out in Kenya and from the USAID Bridges Project administered by Harvard University for fieldwork in Burundi. Professor John Schwille at Michigan State University commented on earlier drafts of this manuscript. His many contributions are gratefully acknowledged.

REFERENCES

- Altbach, P. (1972) *The University in Transition*. Sindhu. Bombay.
- Ashby, E. (1966) *Universities, British, Indian and African*. Harvard University, Cambridge, MA, U.S.A.
- Basu, A. (1974) *The Growth of Education and Political Development in India, 1898-1920*. Oxford University Press, Delhi.
- Berry, J. W. (1985) Learning mathematics in a second language: some cross-cultural issues. *Learning of Mathematics* 5, 18-23.
- Bude, U. (1985) *Primary Schools, Local Community and Development in Africa*. Nomos Verlagsgesellschaft. Baden-Baden.
- Eisemon, T. O. and Nyamete. A. (1990) School literacy and agricultural modernization in Kenya. *Comparative Education Review*, 34.
- Eisemon, T. O. (1989) The impact of primary schooling on agricultural thinking and practices in Burundi and Kenya. *Studies in Science Education*, 17.
- Eisemon, T. O. (1989) Testing Practical Knowledge and Skills. Paper presented to World Bank Seminar on Using Examinations to Improve Educational Quality. Kathmandu. Nepal, 1-3 November.
- Eisemon, T. O., Schwillie. J. and Prouty, R. (1989) Empirical Results and Conventional Wisdom: Strategies for Increasing Primary School Effectiveness in Burundi. Paper presented to World Bank Seminar on Effective Schools in Developing Countries. September 1989.
- Eisemon. T. O., Schwillie. J. and Prouty. R. (1989) Does Schooling Make a Better Farmer? Schooling and Agricultural Productivity in Burundi. Bridges Project Research Report. Harvard Graduate School of Education. Cambridge. MA. U.S.A. (in preparation).
- Eisemon. T. O. (1988) *Benefiting From Basic Education. School Quality and Functional Literacy" in Kenya*. Pergamon. Oxford.
- Eisemon. T. O. Patel. V. and Abagi. J. (1988) Read these instructions carefully: examination reform and improving health education in Kenya. *International Journal of Educational Development* 8. 55-66.
- Eisemon, T. O., Ongesa, E. and Hart, L. (1988) Schooling for self-employment: the acquisition of craft production skills in Kenya. *International Journal of Educational Development* 8, 271-278.
- Eisemon, T. O., Patel, V. and Ole Sena, S. (1987) Use of informal and formal knowledge in comprehension of instructions for oral rehydration therapy in Kenya. *Social Science and Medicine* 25, 1225-1234.
- Eisemon, T. O., Eshiwani, G. and Rajwani, F. (1986) Socio-economic consequences of school expansion and Kenya. *International Journal of Comparative Education* 1,99-137.
- Eisemon, T. O., Hallett, M. and Mandu, J. (1986) Folk tales and school literature in Kenya: what makes a children's story African? *Comparative Education Review* 30, 232-247.
- Eisemon, T. O. (1982) *The Science Profession in the Third World: Studies from India and Kenya*. Praeger, New York.
- Eshiwani, G. S. (1986) Utilization of Examinations. Unpublished paper. Bureau of Educational

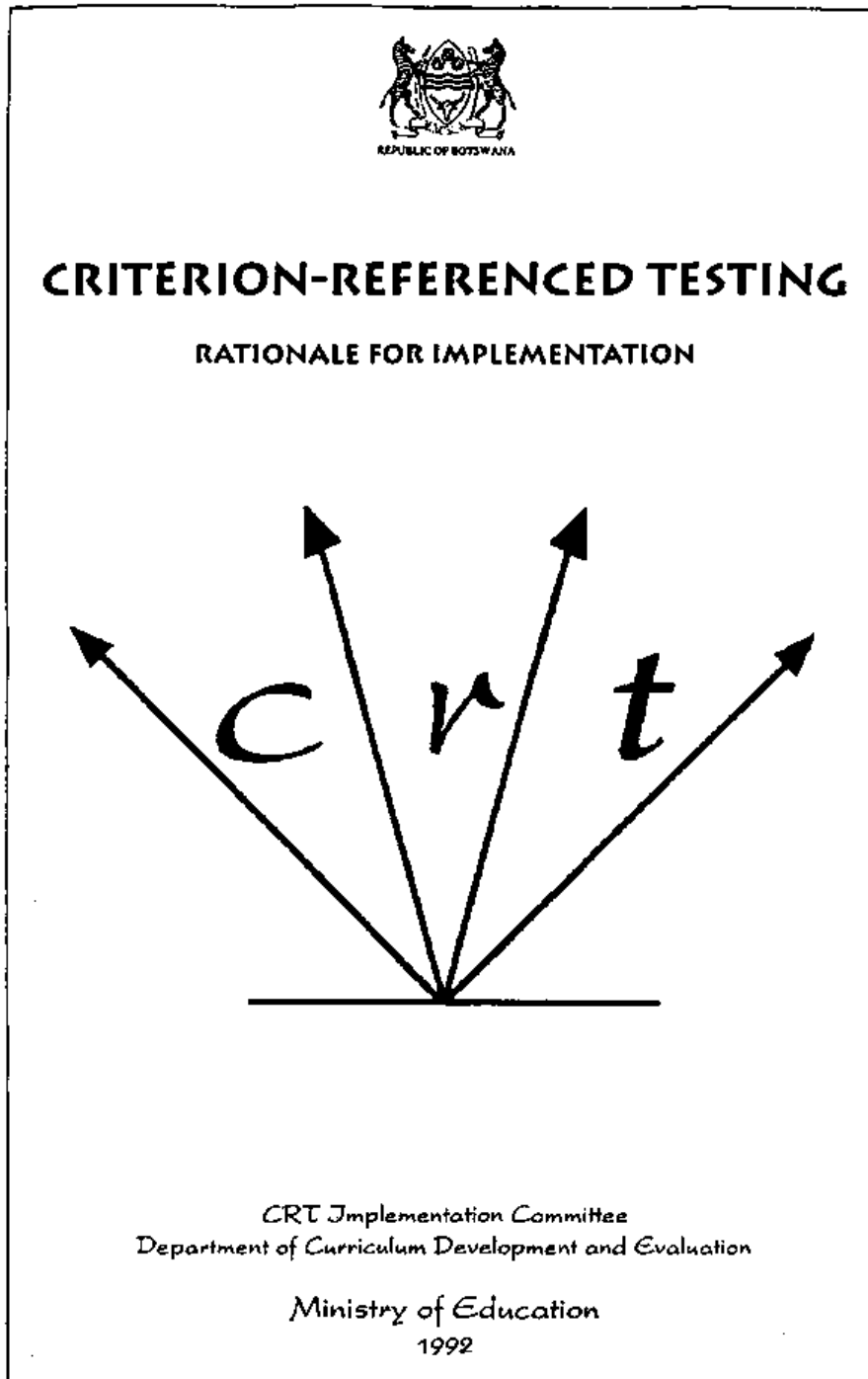
- Research, Kenyatta University, Kenya. Foster, P. (1965) *Education and Social Change in Ghana*. University of Chicago, Chicago.
- Heyneman, S. P. (1987) Use of examinations in developing countries: selection, research and education sector management. *International Journal of Educational Development* 7, 251-263.
- Heyneman, S. P. (1986) Educational Testing to Maximise Economic Performance. Unpublished paper presented at the Annual Meeting of the Comparative and International Education Society, Toronto.
- Heyneman, S. P., Farrel, J. P. and Spuluveda-Stuardo, M. A. (1981) Textbooks and achievement: what we know. *Journal of Curriculum Studies* 13, 227-246.
- Heyneman, S. P. and Loxley, W. (1983) The effect of primary-school quality on academic achievement across 29 high and low-income countries. *American Journal of Sociology* 88,1162-1194.
- Kenya National Examinations Council (1984) *KCPE Sample Papers*. KNEC, Nairobi.
- Ramsay, J. O. and Abrahamowicz, M. (1989) Binomial regression with monotone splines: a psychometric application. Department of Psychology, McGill University (in preparation).
- Savage, M. (1985) The Role of Examinations in Science Education in Kenyan Primary Schools. Kenya National Examinations Council, Nairobi (in preparation).
- Somerset, H. C. A. (1983) Examinations Reform: The Kenya Experience. Unpublished report prepared for the World Bank. Revised and published (1987) as *Examination Reform in Kenya*, World Bank Discussion Paper, Education and Training Series.
- Yoloye, E. A. (1985) Dependence and interdependence in education: two case-studies from Africa. *Prospects* 15, 239-250.
- Zepp, R. A. (1982) Bilinguals understanding of logical connectives in English and Sesotho. *Educational Studies in Mathematics* 13, 205-221.

4.2. Criterion-Referenced Testing. Rationale for Implementation (Abstracts)

CRT Implementation Committee
Department of Curriculum Development and Evaluation

Ministry of Education
1992

(CRT Implementation Committee. Department of Curriculum Development and Evaluation. Ministry of Education, Republic of Botswana, 1992, pp. 3-12)



PRESENT ASSESSMENT PROCEDURES AND PRACTICES

For the Nine-Year Basic Education Programme, the assessment of students through national examinations occurs at the end of each phase in the structure, that is 7 years of primary, the Primary School Leaving Examination (PSLE) and at the completion of 2 years of junior secondary, the Junior Certificate Examination (JCE). The purpose has been to provide information primarily for selection. The question we have been asking each year was 'Who obtained scores falling in the top 20%, 30%, 75% etc. of all the examinees?' The system that best suits this requirement is referred to as 'norm-referenced' testing. The norm here is the specific group of students writing a national examination in any one year. The performance of each student is compared to that of the norm group. A student score can be in the top 10% or 30% relative to the scores of other students taking the same examination.

A similar practice prevails in the school where monthly or end of year examination scores are rank ordered in each class or sometimes stream, and a student is reported as in position 1, 2 or 3. The interpretation that some parents put on the rank order is that their child is the best in the class (which is correct), and therefore that their child performs at a level that should get them an "A" in the external examination (an incorrect conclusion). Because the score only gives a position which is relative to the group that is taking the test, in this case their class, this position can never be generalised to a different norm group, say the national population of the PSLE examinees. When a group changes, the performance pattern may change, so would the relative position of any single score.

The PSLE uses instruments developed by the Department of Curriculum Development and Evaluation (CD&E). There are seven achievement test papers presented for the PSLE. one objective paper for each of the examinable subjects, i.e. Setswana, English, Mathematics, Science and Social Studies; and a composition and letter writing paper in Setswana and English. The PSLE is a norm-referenced test, and this means therefore that the scores of each individual student obtain meaning by being compared to the scores of the other students who sat for the PSLE in the same year. The treatment of scores obtained in the JCE is the same as that of the PSLE.

Limitations of the Norm-Referenced System as Practised

1. *The curriculum coverage of the papers may not remain consistent across the years.*

Curriculum coverage in the PSLE achievement papers is controlled through conscious effort by examiners to have a fair representation of the different units in the final paper. One obvious problem has been keeping this representation of units in the examination paper constant from year to year. Table 1. and Table 2. show the differences in the scheme of questions appearing in the Setswana papers for 1988 and 1989 and the Mathematics papers for 1990 and 1991 respectively.

Table 1. Distribution of items over language categories in the 1988 and 1989 Setswana PSLE Multiple-choice Exam.

<i>Language Category</i>	1988 Items		1989 Items	
	<i>Num.</i>	<i>Percent</i>	<i>Num.</i>	<i>Percent</i>
1. Grammar/syntax	8	13	11	18
2. Word knowledge/vocabulary	27	45	11	18
3. Sentence mechanics	6	10	8	13
4. Use of figurative language	4	7	8	13
5. Reading Comprehension	15	25	21	35
6. Unclassified	0	0	1	2
TOTAL	60	100	60	99

- NOTE:
- a. This table is based on an item task analysis compiled by Naledi Ratsoma, Senior Curriculum Development Officer, Setswana.
 - b. For a classification of the sub-areas of the item task analysis, refer to Appendix 1.

Table 2. Distribution of items over topics in the 1990 and 1991 Mathematics PSLE Multiple-choice Exam.

<i>Mathematics Topics</i>	1990 Items		1991 Items	
	<i>Num</i>	<i>Percent</i>	<i>Num</i>	<i>Percent</i>
1. Sets	3	5	4	1
2. Numbers and Operations	9	15	8	13
3. Fractions and Decimals	8	13	8	13
4. Money	3	5	4	7
5. Time	5	8	8	13
6. Measurement	7	12	9	15
7.. Geometry	15	25	9	15
8. Pictorial Representation	7	12	8	13
9. Algebra	3	5	2	3
TOTAL	60	100	60	99

In view of these differences in the curriculum coverage of the two examination papers, it can be argued that students writing the two papers may have not necessarily been exposed to the same tasks. These two papers are therefore not parallel. If two papers are not parallel, comparison of student performance across the years is definitely limited: hence changes in the standards of performance within each subject are difficult to monitor. This also presents a problem if one wants to compare particular domains or sub-areas in the subject, say 'use of figurative language' as in the case of Setswana. or Geometry in Mathematics. It should however be noted that the problem illustrated here is not only peculiar to the two subjects.

2. *Test results appear to be insensitive to improvements in educational input.*

Test raw scores in each of the five examinable PSLE subject areas are standardised to a common scale (T-scores with a mean of 50 and a standard deviation of 10). The letter grades for each subject are derived from fixed cut-off values, and these have remained constant across years. T-scores are then averaged across subjects and the average score (aggregate) is used for selection. Table 3. shows the T-score cut-off points for the letter grades while Figure 1. is a representation of those on a normal curve.

Table 3. Standard scores (T-scores) cut-off points for average grade levels A B C D and E.

Letter Grade	Standard Score Range	Standard Deviation Range	Percentage of Candidates (approx.)
A	63+	+ 1.3 SD and over	7%
B	55-62	+0.5 SD to + 1.2 SD	24%
C	46-54	-0.4 SD to +0.4 SD	41%
D	-45	Below -0.5 SD	28%

NOTE: The statistics given for the approximate percentages of candidates that obtain different grades each year are averages calculated over four consecutive years.

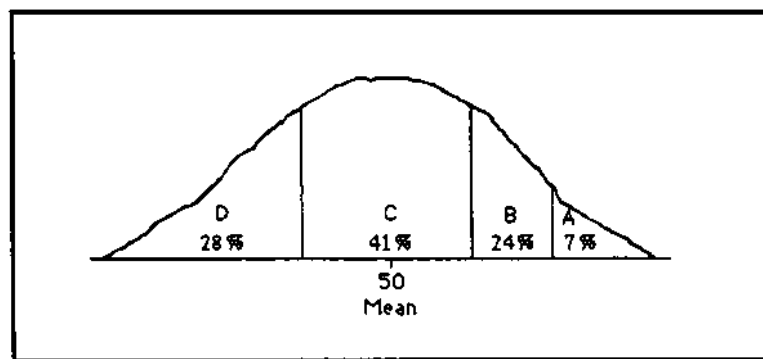


Figure 1

The meaning of each letter grade would remain constant only if the performance of the PSLE candidates remains consistent across the years. In effect what really happens across the years is that irrespective of the changing groups of candidates taking the PSLE nationally, performance on this examination remains consistent because the 'A', 'B', 'C', and 'D' proportions remain the same. These grades therefore, do not tell us much about actual student performances from year to year. They present a rank-ordering of students. In his discussion on selection examinations and achievement testing in Botswana, Somerset emphasised the importance of a more informative assessment system that goes beyond sorting students into a performance order. He alludes to the fact that "Education Officers, teacher trainers, in-service teams, curriculum developers and the teachers themselves should be told which subjects and which topics caused candidates most difficulties so that appropriate remedial action can be taken" (Somerset, 1977).

3. Examination reports do not describe the knowledge and skills which students have.

The current reporting system does not focus on the performance of an individual student with regard to the knowledge and skills instruction was aimed at. Rather reporting focuses on the position of the student in relation to other students who sat the same examination.

4. Test results cannot be used easily as a basis for advising students for career development.

In addition to aptitude tests which provide information about a student's potential ability in different fields of study, career guidance officers need achievement test results to advise students since they (achievement tests) describe student competencies. The present reports on achievement tests do not reflect such competencies, hence they have limitations if they are to be used as a basis for individual career counselling.

5. The correspondence between stated curriculum objectives and the questions which appear on any year's test is usually contested by many hers.

As it has already been established earlier in the discussion, the achievement domains which the examinations sample across the years are quite inconsistent. This is because of the norm-referenced testing procedures, which select test items on the basis of specified statistical features. The items which meet such specifications may not necessarily be those that test the skills emphasised in the instructional programme being examined. As a result, teachers sometimes query the curriculum validity of the examinations.

6. The reliance on a single 'high stakes' examination score (while ignoring many years of student performance in the classroom) is disconcerting.

The PSLE and some subjects areas in the JCE do not use scores from continuous assessments for certification. Teachers have voiced their dissatisfaction with a system of certification that ignores teacher assessments with the argument that they (teachers) have a better opportunity to give more valid assessments of student performances across the curriculum, than a single 1 hour examination paper.

7. The norm-referenced reporting system does not provide the description of individual student competencies in practical subject areas.

The Junior Certificate Examination assesses students in the different practical subject fields, and as in the PSLE, reports reflect a comparison of an individual student performance with the performance of other students. The report does not describe to the prospective trainer or employer the kinds of skills that a student can or cannot perform.

From the above-listed limitations we can conclude that it is difficult to use norm-referenced tests to determine if a student has mastered the specific skills prescribed by their instructional programme, as the primary function of the tests is to provide information about the student's relative standing. This however does not mean that the test data are irrelevant for instructional planning.

AN ASSESSMENT SYSTEM THAT ADDRESSES THE KEY LIMITATIONS OF CURRENT ASSESSMENT PRACTICES.

The preceding discussion presents indications that the present assessment procedures are not compatible with the changes in the curriculum. It is the expressed interest of the current curriculum to focus instructional activities on the learner. An alternative system that may be used to interpret the meaning of students' scores is 'criterion-referenced' approach.

Criterion-Referenced Testing

The early stages of criterion-referenced testing are very closely associated with learner-centred approaches such as individualised instruction or programmed learning, and are therefore designed to measure the mastery of specific learning outcomes or to describe "what Johnny can do" (Popham, 1978). This system is usually used in a situation that requires information on the learning of individual students. The learning of an individual student is in this case compared against the desired learning target. The judgement of individual student performance is against defined criteria, in the form of sets of learning outcomes.

In order to correct the limitations of the present system and retain selection procedures that will continue to be used at national level, a move should be made towards a system whose main function would be to:-

- 1) describe what students can do throughout their schooling, in order to report their achievements as well as diagnose their problems, and
- 2) report their competencies at the end of primary or secondary level education such that grade levels for certification examinations are referenced to a criteria of stipulated competencies.

At school level, assessment needs to be more integrated with the instructional activities, such that it informs the instructional decision making process. Assessment should become a systematic process of gathering information about individual student performance used to describe 'what Johnny, Tshepo, Mpho.... can do'. The main focus of assessment in this mode is to help students learn. This distinction is very vital in the conceptualisation of an assessment system that is programmed to provide feedback to teachers, such that they can make professional judgements about appropriate adjustments in the instructional environment. Testing then becomes a systematic process of gathering evidence of what the child can do relative to the instructional outcome, gathered at an opportune time that will allow the teacher to use the feedback to make judgements about appropriate adjustments in the instructional environment.

Whilst the selection role of the national Primary School Leaving Examination may remain, this will become a secondary function. The main focus of national examinations will be to describe student performances as measured by tests that are fully controlled by the curriculum, with certification criteria that reflect student competencies rather than their relative standing when compared to the norm group.

Key Benefits of Criterion-Referenced Testing

1. Improving the match between the notional examinations and classroom instruction.

"The curriculum that is actually implemented by teachers in schools is, called the **effective curriculum**. The effective curriculum consists of those topics and learning objectives that teachers actually teach to students. The effective curriculum stands in contrast to the **desired curriculum**. The desired curriculum consists of those topics and learning objectives which are found in the national curriculum and which government desires to be taught. The effective and the desired curricula are often overlapping but are also different. Obviously, when a national curriculum exists and when government invests in educational inputs, it is desirable to have these two curricula be congruent." Nitko (1989)

A criterion-referenced test development process provides a common point of reference for instruction, classroom assessment, and external assessments (national examinations). This provision is in the detailed curriculum objectives that precede lesson preparation and test construction. The test blueprint defines desired student learning outcomes; the content or skill as well as the cognitive processes involved. The blueprint presented in Figure 2. attempts to focus both teaching and testing on the same curriculum.

BLOOM'S TAXONOMY	KNOWLEDGE	COMPREHENSION	APPLICATION	ANALYSIS	SYNTHESIS	EVALUATION
READING		1. read a passage and answer questions that follow it.	1. pronounce each word according to spelling	1. identify in text correct punctuation and capitalisation	1. read passage and dramatize it.	
		2. read a text and carry out instructions		2. read text and state its main points.		

- a. This table is based on the STD 7 reading instructional objectives (Setswana).
- b. Every one of these statements is preceded by the phrase Pupils will be able to...
- c. For an example of a complete test blueprint, refer to Appendix 2.

"In the presence of a high-stakes' examination, the key to making the effective curriculum correspond more closely to the desired curriculum is to create examinations that are very tightly aligned with the desired curriculum. The examination must clearly emphasize the student performances which are emphasized by the curriculum. Further, all of the important curriculum objectives should be represented in the test specifications (blueprint) and teachers should be convinced that these objectives may appear on an examination. As a result, the force that motivates teachers to teach to the test is harnessed and directed to the desired end: **Teaching to the test is essentially teaching the desired curriculum.**" Nitko (1989)

Because the 'high stakes' examinations may have such a strong influence on what is taught in the classrooms, an attempt to focus the two curricula on the same area should make teaching more effective and testing more valid.

2. Examinations reports describe the knowledge and skills which students have.

"Since test results can be referenced to pupil performances, one may analyze the results of the national examination to describe what students are capable of doing." Nitko (1989)

A criterion referenced testing system bases instruction and testing on well defined student learning outcomes, which are tasks that students will perform if curriculum objectives have been attained. A test that checks for the attainment of these objectives gives a score that can be interpreted as an indicator of the extent to which a student has achieved the stated curriculum objectives. Controlled sampling of the curriculum allows for the generalisation of a score on a test to performance on the overall domain being sampled.

3. The correspondence between stated curriculum objectives and the questions which appear on any year's test is clear to many teachers.

There are procedures in criterion-referenced testing that enhance the correspondence between a test item (or assessment task) and the objective it is intended to assess. The objective is the starting point. Further definition of the domain described by the objective is given by the cognitive category of the objective in the blueprint. Item writing rules provide further definition of the task through sample items and guide-lines on the construction of the different components of the item.

4. The curriculum coverage of the papers is controlled for consistency across the years.

Table 4. shows an example of a test plan with a representation of the different parts of the curriculum in a test. National examinations based on criterion-referenced test development procedures will have test plans which will serve as sampling procedures. Test plans control the selection of objectives to be tested, and if the same test plans are used in any two years, then curriculum coverage of the two examinations should be comparable.

Table 4 The Test Plan.

STANDARD 5	NO. OF OBJECTIVES	NO. OF ITEMS
<i>Our Country</i>		
Location	5	2
Our Resources	8	3
Culture	2	1
TOTALS		

- a. This table is based on the STD 5 instructional objectives in Social Studies.
- b. For an example of a more detailed test plan refer to Appendix 3.

5. *National educational progress can be evaluated.*

“Since test results can be referenced to pupil performances, one may analyze the results of the national examination to describe what students are capable of doing. Since test specifications (blueprints) remain constant over several years, one may monitor progress on specific curriculum objectives by comparing over the years the percentage of the nation that has learned each objective. Pupils’ performance on clusters of objectives (e.g., those dealing with knowledge of concepts versus those dealing with solving unfamiliar problems) can be compared as well.” Nitko (1989)

Figure 3. represents a situation where national test results do reflect the changing performance of students across time. The information allows for detection of the effects of educational policies, curriculum materials and teaching strategies.

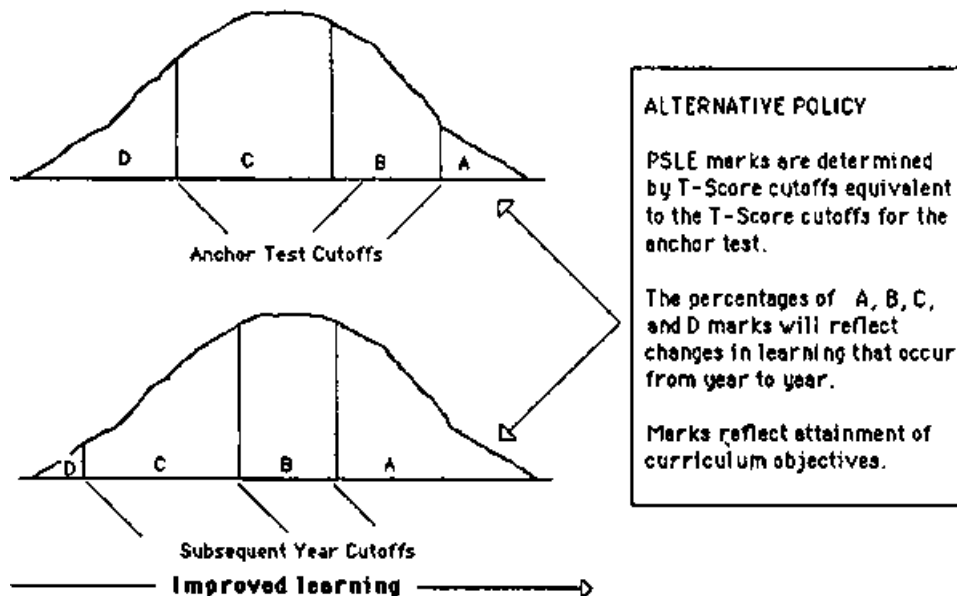


Figure 3

ALTERNATIVE POLICY

PSLE marks are determined by T-Score cutoffs equivalent to the T-Score cutoffs for the anchor test.

The percentages of A, B, C, and D marks will reflect changes in learning that occur from year to year.

Marks reflect attainment of curriculum objectives.

Student specific test results can be used to describe what students in a school or district can do. School progress can be monitored. The information obtained would allow school administrators to make appropriate decisions on improving the performance of the school. In-service programmes would also be targeted on deficiencies revealed by patterns of student achievement in their constituencies.

4.3. Evaluation of Examination Needs of Primary & Secondary Schools in Namibia (Abstracts)

*University of Cambridge Local Examinations Syndicate*¹

¹ International Consultancy & Training Services, October 1990, abstracts pp. 23-34.

Analysis of Question Paper Quality

The following analysis of question papers is an attempt to indicate areas which are likely to need attention when assessment schemes are developed to support curriculum development at Primary, Junior Secondary and Senior Secondary levels....

Physical Appearance

Given that we are dealing with an examination system in which students, in general, do very badly, it is important that the physical appearance of the paper and its layout should be as user-friendly as possible. Quite simply, the students have enough difficulty answering the questions without being placed under greater pressure by the way in which they are presented. This is partially the responsibility of the Examiners who set the paper but there should be quality control systems within the Examinations Division to ensure that papers are properly edited. For example, where a question requires the students to use a piece of stimulus material to answer several part-questions, it is important that the stimulus material can be seen as the questions are read. It is not desirable for students to have to turn over the page as happens frequently at present.

The presentation of question papers would be greatly improved if moderators and Examiners were involved in question paper design and if professional staff of the Examinations Division had a greater editorial role.

This has probably not developed in the past since many papers were directly imported from the Cape but it is a matter which will need attention in the future....

It is also noticeable that the question papers are, in general, lacking in graphic material such as photographs, line diagrams, graphs, charts and maps. This is a matter which will need urgent attention if the question papers are to test higher level skills rather than recall and if they are to be made more accessible to a wider ability range. This would be an important issue in any country but in Namibia where command of the language of instruction is, and will continue to be a problem for the foreseeable future, the use of graphic information will help those whose reading comprehension is weak.

In addition, the use of graphic materials also raises the content validity of question papers in subjects where, in the real world, graphic material is the normal way of communicating information. For example, in a recent Std. 8 English paper, students were asked to extract information from some advertisements but these were presented as plain typescript without artwork and as a result did not look like real advertisements. This shows a technical weakness and is unlikely to interest students already struggling with the language.

It is recommended that Examiners are actively encouraged to consider the physical appearance of question papers and to use graphic stimulus material where appropriate.

It was also observed that many question papers required students to write their answers on separate answer booklets. This required the rubrics to be more complicated than would have been the case had students answered on the question paper. It also causes, as evidenced by comments in Examiners' reports, students to make mistakes in copying question numbers, writing their names on the front etc... Such administrative complexities are confusing for the students but equally importantly they are frustrating for markers trying to locate answers and this necessarily results in a reduction in examiner reliability.

It is recommended that where possible and where economically viable students should answer in spaces provided in the question paper. The layout of the paper should assist rather than hinder students.

Wording of Question Papers

Since Namibia is a country where a very large percentage of the student population have difficulties with the language in which the examination is set, whether it be English or Afrikaans, it is extremely important that the reading level of both the rubrics and the questions is set appropriately. If the reading level is inaccessibly high then the validity of the papers is reduced since the knowledge and skills which the paper purports to test are hidden by the language element. For example, the Std. 4 History paper included the following words and phrases: »erected according to«, »vagrancy of the indigenous people«, »which culminated in«, »led a punitive expeditions« signed in the vicinity« of.

It is recommended that Examiners and moderators are directed to pay particular attention to the reading levels in order to make them more accessible to the target population.

Question Types and Skills Tested

To criticise the content of past question papers is somewhat unfair since Examiners have been constrained by syllabuses which place much emphasis on recall of facts, text book proofs, etc., and, in the absence of training, have closely followed the general format of papers set in the Cape.

A Gazette¹ of the Examinations Board sets out the weightings of different skills to be tested as:

»... every question paper shall be graded to test the pupils' knowledge and skill at the levels of factual knowledge, understanding and insight, in the proportion of 50%, 33 1/3%, 16 2/3% respectively;«

¹ Gazette No. 1/89, Volume 6, 1 March 1989, Examinations Board of South West Africa.

No-one believes that the teachers can follow these instructions and it is generally accepted that question papers set in schools test little more than recall. Moreover, a brief survey reveals that external exams suffer from the same problem. This can be illustrated by an analysis of the Std. 4, Ordinary Grade, History paper for 1989 (Appendix I):

Standard 4 History (Ordinary Grade) 1989

Section A

Questions 1 to 6 were short answer questions all requiring recall of a name. For example, »Name the German officer who managed to free Omaruru from the Herero attack in 1904«, »Name the British commander during the Battle of >Congella< «.

Questions 7 to 12 were a badly set out form of multiple-choice with a negative stem: »Write... the answer between brackets that does not suit the introductory sentence.«

The Republic of Orange Free State (was established in 1854/had Josias Hoffman as its first president/was north of the Vaal-rivier/had Bloemfontein as its capital).

Questions 13 to 18 were short answer questions of the gap filling type. Five of these required a name to be inserted.

»In 1874 Lord Carnarvon sent....., a British historian to South Africa to propagate the idea of a federation.«

Questions 19 to 24 required students to match items in two parallel lists. This is effectively a multiple-choice situation with decreasing numbers of options. The questions all concerned the matching of people's names with an associated act.

Questions 25 to 29 used a map as stimulus material but each question merely required recall of a name or a date.

»When (month and year) was the town numbered 5 on the map conquered by the Union troops?«

Questions in section B required students to write paragraphs containing »at least five facts« on a number of topics.

»Describe the events since 1954 which culminated in the founding of the Republic of South Africa.«

Questions in section C asked for essay-type answers each containing »at least ten applicable facts«.

»Write an essay on the efforts of President M W Pretorius to unite the Orange Free State and the Transvaal.«

As can be seen from the above the question paper places a very heavy emphasis on the recall of names and events. This is a direct result of the teaching syllabus and subsequently ensures that teachers will continue to teach by rote....

The nature of the questions also leads one to question the relevance of what is being tested and hence taught. Further evidence of this can be gained by looking at the Std. 8 paper for Mathematics. However this should be considered with the knowledge that there are serious problems in the teaching of Mathematics in Namibia. Students, teachers and examiners expect extremely low levels of achievement in Mathematics and they are not disappointed in this expectation. It was suggested by a Senior Officer within the Ministry that over the years the students had developed a »phobia for Mathematics«. This was denied by a Mathematics teacher who went on to suggest that it was the teachers who had the phobia! Whatever the truth there is clearly a severe problem which has taken on the form of a vicious circle: poor and unconfident teachers create a fear of Mathematics in students who go on to become unconfident teachers... and so on. This matter clearly requires the concerted efforts of curriculum developers, material writers and teacher trainers. **The assessment system must also play its part by placing greater emphasis on applications of Mathematics in real life situations and at more realistic levels....**

Examination Patterns for the Future

At present the examinations which take place in the primary phase are for the purpose of selection for the next Grade. Whether by design or by accident they have become a harsh filter with the majority of students unable to pass particularly in subjects such as Mathematics and Languages....

With the introduction of automatic promotion in the primary phase the need for a selective examination will be greatly diminished. Indeed prior to the end of primary examination proposed below, the main purposes should be for diagnosis of individual learner difficulties, evaluation of the teaching programme, and reporting student progress at the level of a school report.

It is recommended that there are no external or semi-external summative examinations during the primary phase prior to the PSLE and that schools are permitted to devise their own assessment schemes for reporting purposes.

If necessary, schools should be given guidance and support in the development of non-threatening formative testing methods. They would be helped in this if the primary curriculum explicitly identified levels of achievement which could be tested in each Grade. These **attainment targets** would then act as benchmarks when constructing and grading tests. It is hoped that an additional advantage to be gained from creating a less formal assessment system, at this level is that the administrative pressures on teachers and principals will be reduced allowing them to concentrate on teaching.

Whilst end-of-Grade testing would not be standardised it would be important to introduce standardised tests in a few key areas for the purposes of diagnoses and evaluation. In particular batteries of standardised tests should be available in the fields of basic numeracy and reading.

The frequency of such testing would depend on the curriculum and the attainment targets set. As a minimum, students should be tested annually in reading and/or oral communication and a record of progress kept. In Arithmetic, testing could take the form of criterion-referenced end-of-unit tests to determine whether individual pupils have mastered a particular concept. The tests should be constructed to encourage positive achievement. As a general rule, one would hope to see about 70% of pupils mastering a concept at a 70-80% success rate in such criterion-referenced tests. This is in direct contrast to the low success rates currently achieved in the promotional tests.

It is strongly recommended that the test results be used for identifying individual students within a school who require remedial assistance; diagnostic testing should be followed by active remedial work where necessary.

It is beyond the scope of this report to specify the precise nature of the primary assessment since the primary curriculum is not yet in place.

However, it is strongly recommended that a small panel of experts in primary teaching, curriculum development and testing be set up immediately to consider requirements in this area. Two of the members should be identified as potential Examination Officers of the NEAA to eventually assume special responsibility for primary testing.

The following should be considered guiding principles by the panel:

1. Testing should be non-threatening and should reward positive achievement in the primary phase.
2. Much of the testing will be school-based and will be conducted by the class teacher.
3. Batteries of standardised tests in certain skill areas will need to be produced centrally by NEAA for use by schools.

4. The tests should allow for diagnoses of learner difficulties and for the evaluation of courses/materials.

5. Administration of the tests should not place unreasonable demands on the human or physical resources of a school.

Primary School Leaving Examination

Since the end of Grade 6 will be the exit point from the education system for many students it will be necessary to establish a formal Primary School Leaving Examination.

Since the curriculum for Grades 1-6 has not yet been developed it is not possible to predict in detail what the format should be. This should be agreed by the »task force« created to develop the curriculum.

It is recommended that the primary assessment panel referred to above be involved in both the design of the primary curriculum and the design of the PSLE.

Since it is important that the PSLE is perceived to be an appropriately demanding, valid and reliable examination, it should be administered by the NEAA.

It should also be a predominantly external examination with the proportion of marks allocated to school-based assessment limited to, say, 20%. (If desired this percentage could be increased for predominantly practical subjects provided that adequate controls could be imposed in order to keep reliability to an acceptable level.)

The examination should be certificated by subject rather than by groups of subject. As argued elsewhere, this would ensure that students were rewarded for positive achievement.

Grading should not be by predetermined cut-off points but should be by expert judgement coupled with statistical evidence. This will help to reduce the fluctuations in pass rate due to variation in the level of difficulty of papers from year to year.

The question papers should be designed to reflect the best aspects of the new primary curriculum. In each subject the questions should test the range of skills and content required by the syllabus and not place undue emphasis on recall of isolated facts. The situations chosen for problem-solving questions and others should be relevant as far as young Namibians are concerned and should avoid unacceptable degrees of cultural bias. These issues should be controlled by having each PSLE question paper moderated by a panel of experts as advocated for higher level examinations.

Whilst the papers should maintain appropriate standards and should include tasks which the more able would find demanding the emphasis should be on allowing students to show what they know and can do. They should therefore be designed to give a higher mean mark and standard deviation than exams do at present. Associated with this, the marking schemes or memoranda should be detailed enough to ensure inter- and intra-examiner reliability but should not punish all errors, however trivial, as tends to be the case at present.

Since it is highly desirable that the PSLE should reflect a new philosophy of both teaching and testing it is recommended that question paper setters are trained in the theory and practice of item writing before specimen papers are developed for the new curriculum.

Junior Secondary School Certificate

It should be noted that a new broad curriculum for the Junior Secondary phase (Grades 7, 8,9) and proposed syllabuses for the subjects included in that curriculum were presented publicly in Namibia in October 1990. Included in the proposals was an outline for assessment procedures during Grades 7 and 8 only.

The following principles were incorporated:

- (a) the assessment would be school-based with no external components;
- (b) in all subjects, continuous assessment would be allowed up to a pre-determined maximum weighting (greater for practically based subjects);
- (c) formal tests would be set and marked by teachers;
- (d) promotional criteria would be based on success in specified subjects or groups of subjects rather than on an aggregate mark.

4.4. Primary Science School Leaving Examination (Abstracts)¹

by Richard Bradbury

¹ Malawi-German Primary Science Project. Occasional Papers No. 1. Domasi. October 1992, pp. 2-8. **Note:** Since 1995 the PSLCE uses multiple-choice test items. Science is now part of a paper called »Science Incorporated« consisting of three parts: Section A »Agriculture« (25 marks), Section B »Science« (30 marks) and Section C »Health Education« (15 marks).

Background to the Primary School Leaving Certificate Examination

Pupils who reach the end of the eight year primary education course in Malawi, in theory at the age of 14 but on average at 15, are expected to take the Primary School Leaving Certificate Examination (PSLCE). This is an external examination set and marked by the Malawi National Examinations Board (MANEB), a parastatal organization which is responsible for all primary and secondary school examinations in the country and for certain other kinds of testing. The PSLCE undoubtedly has a great influence on what is studied in the schools and on how it is taught, particularly in the latter half of primary education (Standards 5 to 8). This is not only because the examination syllabuses (largely perceived through past examination papers) become the de facto curriculum, but also because the PSLCE acts as a selection test for admission to secondary education. Since there is very limited provision of secondary places (under 10 per cent over 100,000 PSLCE candidates each year are accepted for secondary school) there is great competition to do well in the examination. Pupils, parents and teachers insist on studying only those things which are known to be tested in the examination. The consequence is that aspects of examined subjects which are not in the examination are ignored and other parts of the curriculum which are not examined at all, such as physical and religious education and art, are simply dropped in the last few years of primary education.

It is therefore necessary for any curriculum development project in primary education to take this examination into consideration. In view of the shortage of secondary school places, it is unlikely that it will be abolished in the near future (this was anyway tried in the early 1960s with apparently disastrous consequences; Banda, 1982). It should instead be regarded as a necessary evil which can with skill and effort be managed as a positive influence on what is taught and learned. In educational jargon, curriculum development in these circumstances should be assessment-led.

There has recently been some attempt to use this approach to influence the curriculum through the PSLCE. For many years, the examination was administered by the Ministry of Education and Culture (MOEC) with only the final-year secondary examination, the Malawi Secondary Certificate of Education (MSCE) being run by an examining Board, then called the Malawi Certificate of Education and Testing Board (MCETB). The Ministry provided little more than basic administrative support for the examination and the professional inputs to ensure quality were lacking. The question papers in almost all subjects regressed to testing simple

recall of often trivial facts, with little or no attempt to test the so-called higher level skills such as comprehension, analysis, reasoning or problem solving. This state of affairs was generally bemoaned but nothing was done about it until 1987 when, with World Bank support, MANEB was formed to replace the MCETB and to take charge of all school examinations in Malawi. A report by Somerset (1987) detailed the deficiencies of the current examinations, particularly PSLCE, and made recommendations for improvements. Between 1989 and 1992 a British test development consultant was attached to MANEB and began to implement some of these recommendations.

A brief summary of what was done may be of use to future curriculum development. The setters and moderators of questions in all subjects were assembled for a one week workshop. Starting from the axiom of »what you test is what you get« (WYTIWYG) they were asked to scrutinise past question papers in order to find out what had been tested in terms of both content and abilities. They were then asked to decide what ought to be tested in the light of the current syllabuses, teachers' guides and pupils' books. Large mismatches were found, particularly in the practically orientated subjects like Agriculture, Home Economics, Needle Craft, and Science and Health Education; the guides and books often promoted an activity or experimental approach to the subjects which was not reflected in the examinations. Participants were then asked to draw up a content/ability specification grid for their subject, showing what balance of each they would wish to test.... Using the grids, they were next asked to set a specimen question paper reflecting this balance. As guidelines, they were given the »principle of the 3Ps«, viz, that questions should as far as possible be based on:

- Problem: pupils should be asked to find an answer from data given or situations shown, rather than just to supply a fact from memory;
- Practical: questions should be based on real situations and realistic contexts and as far as possible have some relevance for people in Malawi;
- Picture: situations should be presented through drawings, photographs, diagrams, maps, graphs or tables as appropriate.

These principles were designed to confront the worst features of the current papers which featured such questions as »name two kinds of latrines«, or, worst of all, »what garment is made in Standard 6«.

Among subsidiary principles used were the idea that a good examination question should be usable as worthwhile teaching material, and that the teacher's guides and pupil's books were a good source for the »3Ps« if their material were adapted somewhat to show slightly unfamiliar situations which could be used to test pupil's understanding of the underlying principles.

The question papers produced by the workshop were later edited and pretested in sample schools to check that the new style of questions was not too difficult for pupils at this level. Following satisfactory results, the papers were printed in quantity and sets sent to each primary school in the country with Standard 8 pupils. Receipt was monitored by means of reply-paid acknowledgement cards. The schools were also informed that papers of this type would be set in the PSLCE from July 1991 onwards, which has indeed been the case.

It may be gathered from the above account that the intention of recent test development has been to exploit the present syllabuses and teaching materials to best advantage, and to encourage teachers by the example of the examinations to widen the range of abilities taught. This has been attempted in all subjects but it is of particular relevance to Science and Health Education where a reasonably good if incomplete set of teachers', guides and pupils' books exists.

It may be worth noting at this point that Somerset (1987) made the following observations:

- »In developing new curricula it is important that teaching and learning goals on the one hand, and assessment goals on the other should be closely co-ordinated. If the

development of prototype assessment instruments is delayed until after other work on the curriculum has been completed, there is a real danger that the profile of skills tested by the assessment instruments will diverge significantly from the profile specified for teaching and learning. The new curriculum should include sections giving examples of questions which could be used by teachers and examiners to measure the progress of pupils towards mastery of the desired competencies. Learning and assessment should always be seen as complementary parts of the same process. «

These points should be borne in mind for future curriculum development which must be based on the new syllabuses produced by the Malawi Institute of Education (MIE) and shortly to be introduced into the schools. These syllabuses show from their columns for »suggested assessments« that Somerset's points have so far been ignored. There is a danger that examinations based on the new syllabuses may prove once again a restrictive influence unless some further test development is begun well before pupils in Standard 5-8 begun to be taught the new material in 1996 - 99...

Science and Girls

Whilst all boys study Science, girls are given the option of studying either Home Economics or Needle Craft instead. Girls are in any case under one third of pupils entered for the PSLCE. In 1992, of 106,073 candidates, 34,368 or 32.4 per cent were girls. Of these girls 11,408 did Home Economics (3,061) or Needle Craft (8,347) being 33.2 per cent of girls and 10.8 per cent of all pupils. Those taking Home Economics are examined on Health Education as part of it, but those taking Needle Craft are not examined on it at all. There is thus some minimal element of Science in Home Economics but none at all in Needle Craft. Girls are already at an educational disadvantage in Malawi as the low proportion enrolled in Standard 8 shows: social and economic pressures force them to drop out of school early. They also perform less well than boys in all subjects of the PSLCE. Lack of Science education will further handicap that minority which succeeds in going on to secondary school where Science is compulsory. The MOEC has considered abolishing this practice of making Science optional for girls, but it seems unlikely that this will be done until the new curriculum takes effect for Standard 5 in 1996...

Nature of Questions Set in PSLCE Science

The Science questions in the paper set in July 1992 have been analysed... The extent to which school science apparatus is shown in them has been noted, as has the occurrence of experiments or procedures described verbally which could have been pictured. It should be noted that all the apparatus shown is simple and made from cheap and easily obtainable materials (except batteries and bulbs). All could be used by a teacher to demonstrate or done by pupils themselves, even at home. In this respect, the setters have observed the 1989 - 91 phase of test development and based questions on material shown in the teachers' guides and pupils' books which should be within the capacity of any teacher to demonstrate in the classroom...

Seventeen questions were set, nearly all having sub-questions with marks separately apportioned. Of these 17 questions, 6 were on biological topics which do not require apparatus other than the organisms studied. Of the remaining 11, 6 showed apparatus and an activity or procedure to be done with it. Of the other 5 questions where a verbal description was used instead of a drawing, it would have been possible to use an illustration instead. Indeed, illustrations have been used for some of these topics in the 1991 paper or in the specimen paper.

Of abilities tested, 10 questions seem to test recall only. Two questions which look as if they test more than this by requiring pupils to reach conclusions from observation of the results of experiments (Q5 on water pressure and Q17 on permeation of soils by water) are in fact exact copies of the experiments in the teachers' guides and could be answerable by recall alone. Of the remaining 7 questions, 5 do appear to test application of principle in that they show or describe procedures or experiments which are slightly different from those in the teaching materials, but which can be understood and interpreted by use of the appropriate

principle. Two questions appear to test comprehension or interpretation.

An allocation of marks to abilities tested shows that out of 60 marks for the Science section of the paper, 35 were given to recall sub-questions, 5 to comprehension and 20 to application of knowledge or principle. This division (which is of course subjective) does not achieve the target of 50 per cent of marks for recall questions and 50 for more than recall... although it would if Q17 had been adapted properly.

It seems reasonable to conclude that the 1992 paper tried to test the specified range of abilities. In terms of content, there were more marks for biological questions (20) than is intended by the specification grid (15). It is again a subjective judgement, but there seems a small tendency to regress to the pre-1991 style of questioning with questions of the »name this...« type. This type of question was never actually abolished, which is a pity since its continuance may encourage a regression to more recall questions...

Support for Testing Higher Level Objectives

It is to be hoped that the above account demonstrates that MANEB's test development policy is to make examination questions support higher level objectives as far as possible. The only limitations are the syllabuses and teaching materials, which in the case of Science are not a serious obstacle, and the need to maintain some continuity with the old form of questioning by retaining some items to test recall. The effect of recent changes of the emphasis of questions on the ways in which teachers use this material or adjust their teaching approach is of course not known, although it might be possible to do some research on the topic. In order to exploit the changes in question style more, MANEB and MOEC ought to make more use of examination data to draw the attention of teachers to the real point of some of the questions and to appropriate ways of teaching for them. This technique has been used in Kenya with apparent success (Somerset, 1987). Unless this is done, the support which the examinations are trying to give to the higher level objectives remains passive. The idea of reporting examination data to teachers and organizing local seminars to discuss it has been raised here, mainly by the World Bank (1989) but seems to have provoked no reaction from MANEB and MOEC.

What Pedagogic Approach to School Science is Reflected in Questions?

From the above, it may be evident that recent test development has tried to encourage an approach to Science which is based on the use and observation of real materials and apparatus, on simple experiments to be performed, and on making conclusions from experiments. A contrary approach before 1991 tended to emphasise naming and definition, the acceptance of principles as facts with no need for experimental demonstration, and minimal use of any inductive or deductive reasoning. This latter approach, although still detectable at times, has been discouraged by the recent reforms. It is now in the teachers' best interests in terms of examination success to ensure that pupils have at least seen the demonstrations and experiments in the teachers' guides, and preferably also those in past question papers. Likewise, pupils should be made aware of the need to predict the results of experiments and explain them, using appropriate principles. If the general view that learning is more successful through »interaction with objects« (Piaget) is accepted, pupils ought to do better as a result of being taught via concrete demonstration and preferably via direct participation. But it is necessary to repeat here that the support of the examinations for this approach to teaching is at best passive. The material and attitudinal obstacles to teachers adopting it in practice are well known (Siege and Voss-Lengnik, 1990) and it is obvious that intervention of some kind will be needed to influence teachers more radically.

It may be thought that the concept of Science promoted by the PSLC syllabus, teaching materials and examination alike is a very limited one, even allowing for the scarcity of resources in Malawi. There is little to encourage pupils to devise experiments of their own or to deduce principles from observation. There is no social context for Science, which must seem a baffling activity to pupils in a rural society. There is no acquaintance with the discoveries of great scientists, which might convey human interest as well as some appreciation of scientific method by empathy. Nor is the use of Science made clear, as it

easily might be in relating, say, discoveries in medicine to daily life. It is not the fault of the examinations alone, but a very isolated and mechanical concept of Science is being promoted in the Malawian curriculum, both old and new.

4.5. Learning Competencies For All. Essential and Desirable Learning Competencies for Standards 4, 5 and 6 (Abstracts)

Towards a holistic approach to examination reform

Prepared by:

Mauritius Examinations Syndicate
in collaboration with
Ministry of Education and Science,
Mauritius Institute of Education
and Mahatma Gandhi Institute.

LEARNING COMPETENCIES: AN INTRODUCTION

'A nation's children are its greatest resource. In only a few decades the prosperity and quality of life of all nations will be determined By to-day's children...'

Cited in: *Master Plan for Education for the year 2000*

Why the Learning Competency Project?

The Philosophical and Sociological Background

In an age where illiteracy has become a major handicap, every nation owes it to her children to make sure that all of them irrespective of class and creed have access to at least the basic education that will enable them to become literate and functional citizens. This is rightly now the international trend.

Mauritius which has always given due importance and status to education has followed suit: universal free education is available at the primary, secondary and tertiary levels. Yet a closer look at the real situation shows that it is not yet time for us to rest on our laurels.

"At present while 99% of our children enter standard I, only a little more than 70% of the cohort pass the CPE examination. One quarter of our children leave school at that point without basic education"

acknowledges Honourable A Parsuramen, Minister of Education and Science. In other words nearly 30% of our children leave school, after six years, illiterate, innumerate and labelled as failures. This wastage of human potential and of economic and human resources has to be tackled urgently. Again to quote the Minister of Education and Science:

"Mauritius needs the talents of all its people. We cannot afford to ignore the potential of large groups of our citizens, nor indeed would it be socially just for us to do so."

Having achieved the first goal of primary education, we have now a social and moral obligation to improve the quality of education to ensure that *all* our children irrespective of social, regional, economic background are given *quality education* and helped to develop their abilities and the basic life skills and competencies necessary to function in the present society.

The Learning Competency Project was initiated by the MES with precisely these objectives in view:

- (1) to identify the basic skills and learning competencies needed by children to become literate functional citizens;
- (2) to provide direction for curriculum developers to develop competency-based instructional materials;
- (3) to provide broad guidelines to teachers to adapt teaching learning strategies to the learning competencies;
- (4) to redesign the examination papers in terms of the learning competencies;
- (5) to provide a basis for certifying pupils' achievement.

Pedagogical Basis

While it is necessary and desirable to get the majority of pupils to pass the CPE examination, one major pedagogical concern is to improve the performance of pupils to bring them all up to a reasonable attainment level - to make them really literate and numerate.

By setting clearer and step-by-step attainment targets (expressed in terms of essential and desirable learning competencies), this document proposes an in-built mechanism for more effective teaching. It gives a clearer sense of direction in the day-to-day teaching so that teachers may know step-by-step what they are expected to achieve and they can discover/diagnose at what step a child is facing difficulties.

The document also provides the basis for *target-related assessment* which can take various forms:

- (a) diagnostic to identify pupils' difficulties;
- (b) formative for improvement;
- (c) summative for certification.

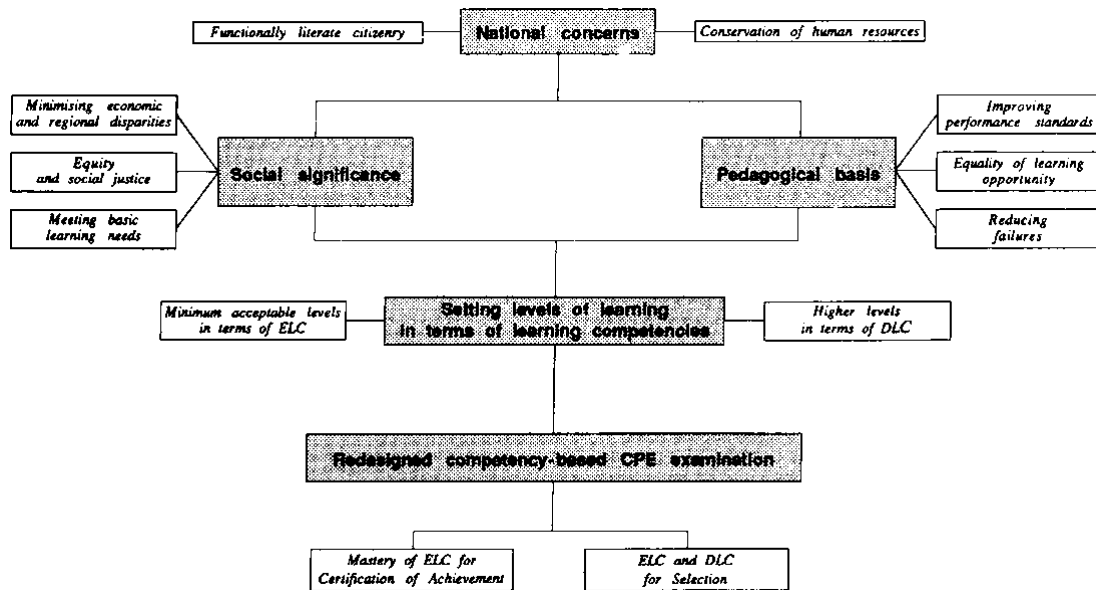
What are Essential and Desirable Competencies?

Essential Learning Competencies (ELC) represent the levels of learning in a particular subject comprising such basic knowledge, understanding, skills, abilities, interests, attitudes and values which are considered minimum but essential for all students to acquire at the end of a particular standard or stage. They can be regarded as attainment targets below which learning is not sustainable. In other words, they are the 'minimum vital'.

However, children do not all have the same potential and while it is necessary to bring all children up to the agreed-upon essential level of learning, children with higher abilities must be catered for and attainment levels must be pitched higher to meet their learning needs. Therefore, higher order competencies involving more complex mental processes and/or learning content have been laid down and termed **Desirable Learning Competencies** (DLC). The ELC are a must for *all* students while DLC are optional though desirable, for every one to exercise his higher order mental faculties and can thus be used to discriminate amongst high flyers.

For examination purposes a judicious combination of the two types of competencies will be used. The ELC are more suitable for certification while DLC are more appropriate for ranking purposes.

The theoretical framework of the Learning Competency Model may be depicted as In the figure on page 3.



Theoretical Framework for teaming Competencies

Methodology used to formulate Learning Competencies

The **first step** in the project was to analyse the present syllabuses and question papers and to study the international literature on competency-based teaching and testing.

Different models have been used in different countries. *Basic Skills Testing Programmes* are used in Australia where the skills tested include two aspects of literacy (Reading and Language) and three aspects of Mathematics (Number, Measurement and Space) and where the scores represent growth along a continuum. In some Australian states, profiles have been developed within subject areas and each component is further divided into levels of competence representing standards of performance.

England and Hong Kong have specified “*Attainment Targets*” with criteria set at a number of different levels rather than pass/fail at only one level. The USA have *Minimum Competency Testing* and India developed the *Minimum Levels of Learning*.

The Models may vary from country to country but they all acknowledge the impossibility of measuring competence, as such. We can only define objectives and measure the *behavioural outcomes* of these objectives. When we teach a lesson to develop a child’s ability to understand - we cannot examine and certify whether the thought processes involved in understanding were developed and used in that particular lesson, we can only assess his level of understanding by the answers he gives to the questions we ask and depending on his answers - which are measurable - we can decide about his level of understanding. Hence the care with which tests (class tests as well as external ones) should be constructed to ensure that they measure what they are supposed to measure.

The MES chose a taxonomic model which states learning objectives in terms of:

Knowledge, Understanding and Application for content subjects and Knowledge, Comprehension, Expression (which includes Application) in the case of languages.

This model is close to the Indian Experience but different and more adapted to the Mauritian context.

The **second step** in the project was to break down each subject into its major skills and the competencies implied in language subjects or content areas and the corresponding competencies. Thus Learning Competencies were laid down for each subject examined at CPE level giving due importance to certain skills presently neglected and overlooked which are yet essential components of the subject, e.g. the oral skills in languages and the psychomotor skills in EVS. We have also considered some of the non-cognitive elements like attitudes, values that are important not only for the development of competencies in individual subjects but also for the healthy growth and integration of the child in society.

The **third step** was the categorisation of the Learning Competencies into the two groups: **Essential and Desirable**. *Essential Learning Competencies* constitute the levels of achievement to be developed in *all* children at the end of the primary stage while *Desirable Learning Competencies* set the attainment levels for children with a *higher ability*.

Strategy for formulating Learning Competencies

To develop the lists of Learning Competencies, the MES adopted a *participative* strategy.

Subject working groups comprising curriculum developers, chief examiners, inspectors, head teachers, deputy head teachers, practising teachers and research officers were constituted to work out the Learning Competencies in each subject.

A Steering Committee was also set up to monitor the progress of the panels.

These were the different stages in the development of the project:

(a) August to November 1991

After a careful analysis of the prescribed textbooks, of question papers and of relevant documents, the subject working groups, under the guidance of Prof Pritam Singh laid down the competencies for each subject. This material was regularly reviewed by the Research and Development Section of the MES in accordance with the observations of the Steering Committee.

(b) 10th to 16th December 1991

The first draft of Learning Competencies was prevalidated by about 250 practising teachers who had been sent the list of competencies in advance. Advantage was taken of their experience (i) to identify the competencies to be deleted, modified or added and (ii) to identify those competencies that can be acquired only by bright students and are not essential. Such competencies - named *Desirable Learning Competencies* are listed in **blue** in the document.

(c) December 1991

To ensure further judgemental validity the modified draft of Learning Competencies was submitted for critical and thorough review by senior experts from NCERT, New Delhi, who have experience in this field.

(d) 22 April 1992: National Seminar on Reforms at CPE Level

The seminar on proposed reforms at CPE level provided some reflections and observations which were incorporated.

(e) July 1992

Final Draft.

The following criteria were applied to judge whether a Learning Competency formulated is acceptable for inclusion in the list.

Firstly, it must be *sustainable* in the sense that it represents achievement which can sustain learning from one unit to the other and from one standard to the next so as students can derive benefit from instruction for further learning.

Secondly, it must be *communicable* which means that the levels of learning stated in the document in the form of learning competencies form a common basis for the teachers, evaluators, inspectors and administrators. Each Learning Competency has a common pattern of statement representing two dimensions, namely the content element and the ability aspect, to facilitate communication.

Thirdly, an attempt was also made to provide *learning continuity* in each topic and sequenced in such a way that clusters of competencies of a unit are built upon the clusters of competencies of the preceding unit. An attempt was also made to develop a continuum of learning competencies as far as possible across standards 4 to 6 besides the learning continuum established within each standard.

Fourthly, the criterion of *functionality* was used. It means that teachers are capable of developing these competencies in teaching. Learning Competencies are stated at a proper level of generality, neither too global to be measurable nor so atomistic as to be unwieldy.

Fifthly, unless a learning competency was *measurable, i.e. evaluable* it was not to be listed. A learning competency must provide a well defined goal, where a statement in terms of specific learning outcome is necessary, to make it testable.

Sixthly, *achievability* was the final criterion which means that under the given conditions all learning competencies are attainable. They are in accordance with the cognitive development and the maturity levels of students.

Why the hierarchical presentation?

A look at the document shows that the competencies are presented in a hierarchical way both across standards and more importantly within each standard. The reason is that learning objectives have an intrinsic hierarchy

knowledge → understanding → application

which is reflected in both content and competencies as a result of which we have a hierarchy of competencies and a hierarchy of content. Logically we cannot expect a child to multiply till he understands the principle of addition; likewise we cannot expect him to write a sentence correctly in a particular language till he has mastered certain syntactical and grammatical structures of the language and has the required vocabulary. This hierarchical nature of the competencies listed has a direct implication for teaching: given their hierarchical nature, a competency cannot be taught unless the preceding one has been acquired. Teachers must therefore make sure - not necessarily through formal tests - that the pupils have acquired the prerequisite competency before they teach the next one.

To Conclude

The present document is not a teaching syllabus nor an examination syllabus. Its main focus is not on the content elements of the different subjects but on the *acquisition of the competencies*. Although there are certain subjects like EVS and Maths which are more content oriented than others and all subjects have a content component, in the present document content is above all a tool for the development of the identified competencies.

This explains the use of **action words** to define all the competencies to be taught. At the end of a particular standard we expect the child to be able to *recall, decode, analyse, interpret etc...* It is immaterial whether the teacher uses passage 1 or 2, topic 1 or 2 to achieve this, as long as the child develops the required competency and reaches the level of *knowledge, understanding* and *application* appropriate for his age. The success of the Learning

Competency Project rests on *competency-based teaching and testing*.

The Learning Competency Project is part of a continuing search for standards in education. As such it relates to a particular time frame: as the demands of society change, the acceptable levels of learning and areas of essential learning have to be reviewed and updated. As performance standard goes up with improved instruction and better learning, the list of ELC also changes and some of the present DLC become ELC.

Notes

1. Learning Competencies have been formulated for Standards 4,5 and 6 for all subjects examined at CPE level. The prerequisites for Standard 4 have also been listed.

2. Code used for the Learning Competencies

The *first* figure represents the skill/topic referred to, the *second* refers to the standard and the *third* to the competency to be developed.

Thus in the code **2.4.6** for English, **2** represents the second skill listed i.e. *speaking*; **4**- *Standard 4* and **6** - the *sixth competency* to be developed.

Topics	Prerequisites	Standard 4	Standard 5	Standard 6
				6.6.12 <i>Acquire and practice good health and safety habits.</i>
				6.6.13 Give reasons for taking a balanced diet.
7. Air		7.4.1 Show the presence of air in the environment.	7.5.1 State that air is a mixture of many gases such as oxygen, carbon dioxide and nitrogen.	
		7.4.2. Demonstrate that soil contains air.	7.5.2 Demonstrate that air is necessary for burning.	
		7.4.3 Demonstrate that water contains air.	7.5.3 Name items which easily catch fire.	
		7.4.4 Name some light objects which can float in the air.	7.5.4 List some of the ways to extinguish fires.	
		7.4.5 Conclude that air is present all round us even in 'empty' space.	7.5.5 List certain precautions to be taken to prevent fires.	
		7.4.6 Demonstrate that air exerts pressure.	7.5.6 Explain three causes of air pollution.	
			7.5.7 State that air is essential for living things.	

Topics	Prerequisites	Standard 4	Standard 5	Standard 6
			7.5.8 Infer that air exists around us.	
			7.5.9 Demonstrate that air expands on heating and contracts on cooling.	
			7.5.10 Give three reasons why air pollution is dangerous.	
			7.5.11 Infer that air exerts pressure in all directions.	
8. Weather	8.1 Describe prevailing weather (from pictures).	8.4.1 Observe and record weather characteristics.	8.5.1 State that wind is air in motion.	
	8.2 Interpret a simple weather calendar.	8.4.2 Name and recognise the instruments used to measure temperature, rainfall and wind direction.	8.5.2 Show that wind has speed and direction.	
	8.3 Record the weather using symbols in a calendar for a week.	8.4.3 Name the winter months and summer months.	8.5.3 Read the temperature from a thermometer.	
	8.4 Record the weather in a weekly calendar, using symbols.	8.4.4 Mention the weather characteristics of summer and of winter months.	8.5.4 Read and recognise a map with isotherm/isobar/isohyet.	
		8.4.5 Name the prevailing winds.	8.5.5 List the three main types of rainfall.	
		8.4.6 Explain the importance of rain.	8.5.6 Measure rainfall using a simple rain gauge.	
		8.4.7 <i>Develop an interest in weather reports on radio and television.</i>	8.5.7 Make models of anemometers and windvanes in small groups.	
		8.4.8 <i>Develop an awareness that our weather is influenced by cyclones in</i>	8.5.8 Differentiate between land breeze and sea breeze.	

Topics	Prerequisites	Standard 4	Standard 5	Standard 6
		<i>summer.</i>		
			8.5.9 Interpret isotherm/isobar/isohyet maps.	
			8.5.10 Differentiate between weather associated with cyclones and anticyclones.	
			8.5.11 Interpret cyclone bulletin.	
			8.5.12 Explain how rainfall distribution in Mauritius is related to relief, seasons and the South East trade winds.	
			8.5.13 Interpret a rainfall histogram.	

4.6. Science in the National Curriculum - Attainment target 9: Earth and Atmosphere

Knowledge and understanding of science, communication and the applications and implications of science (ATs2-17)

Pupils should develop their knowledge and understanding of the structure and main features of the Earth, the atmosphere and their changes over time.

LEVEL STATEMENTS OF ATTAINMENT

Pupils should:

- 1
 - know that there is a variety of weather conditions.
 - be able to describe changes in the weather.
- 2
 - know that there are patterns in the weather which are related to seasonal changes.
 - know that the weather has a powerful effect on people's lives.
 - be able to record the weather over a period of time, in words, drawings and charts or other forms of communication.
 - be able to sort natural materials into broad groups according to observable features.
- 3
 - be able to describe from their observations some of the effects of weathering on buildings and on the landscape.
 - know that air is all around us.
 - understand how weathering of rocks leads to the formation of different types of soil.
 - be able to give an account of an investigation of some natural material (rock or soil).
 - be able to understand and interpret common meteorological symbols as used in the media.
- 4
 - be able to measure temperature, rainfall, wind speed and direction; be able to explain that wind is air in motion.

- know that climate determines the success of agriculture and understand the impact of occasional catastrophic events.
- 5
 - know that landscapes are formed by a number of agents including Earth movements, weathering, erosion and deposition, and that these act over different time scales.
 - be able to explain how earthquakes and volcanoes are associated with the formation of landforms.
 - be able to explain the water cycle.
 - 6
 - be able to explain the processes by which igneous, sedimentary and metamorphic rocks were formed and are recycled.
 - be able to describe how the properties of minerals and rocks are related to their uses as raw materials.
 - understand how different airstreams give different weather.
 - 7
 - be able to state qualitatively the relationship between pressure and winds.
 - be able to recognise patterns in the distribution of the Earth's major surface features (continents, mountain belts, areas of very old rock, oceans, ocean basins, trenches and ridges) and zones of active crust (earthquakes and volcanoes).
 - 8
 - understand that geological time scales are very long compared with human and historical time scales, and have a general knowledge of how geological time scales can be measured.
 - be able to interpret evidence of modes of formation and deformation of rocks.
 - 9
 - be able to use appropriate scientific ideas to explain how changes in the atmosphere cause various weather phenomena.
 - be able to describe in simple terms the layered structure of the inner Earth, and explain the evidence that favours such a model.
 - 10
 - understand the theory of plate tectonics and use it to explain some major geological features on the Earth's surface.
 - understand how plate tectonic theory brought about a revolution in our understanding of the way the outer part of the solid Earth works.

Source: Department of Education and Science (1991)
Science in the National Curriculum. HMSO. London.

4.7. Boat Building (The Properties, Classification and Structure of Materials) - Worksheet

Statement of attainment

Pupils should:

3/3a be able to link the use of common materials to their simple properties.

Resources

- Copies of the worksheet for this activity

Acceptable responses

A choice of material for each part of the model should be made. The actual materials chosen are less important than the reasons given for choosing them. In total, at least two different properties of materials should be named or described.

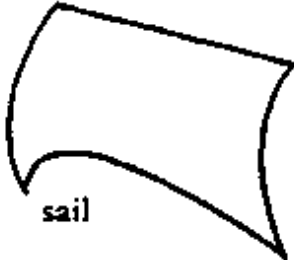
Suitable properties would include:

- strength
- weight (density)
- elasticity
- hardness
- flexibility
- non-absorbency

The correct terms need not be used; it is sufficient to convey the meaning of the property (for example, 'it won't tear' or 'it would be easy to cut').

Boat Building

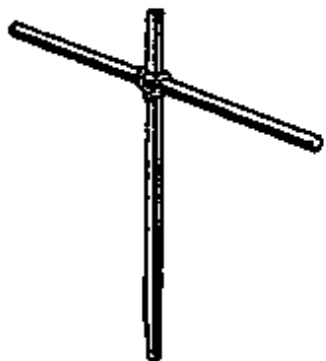
Example answers only - others are possible



I would use paper / nylon fabric / foil

because it is strong


and catches the wind



I would use rubber / card / wood

because it is stiff

and light



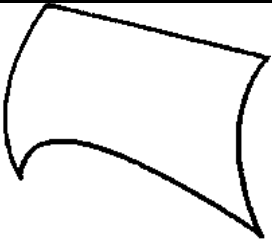
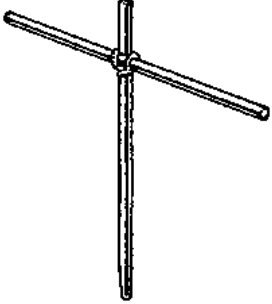
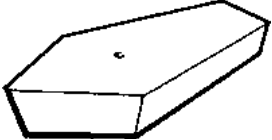
I would use wood / polystyrene / plasticine

because it is light or floats

and easy to cut

Materials - Levels 2-5

Boat Building

 sail	I would use because it is and	paper/nylon fabric/foil <hr/> <hr/>
 mast	I would use because it is and	rubber/card/wood <hr/> <hr/>
 hull	I would use because it is and	wood/polystyrene/plasticine <hr/> <hr/>

© National Foundation for Educational Research, 1992

This sheet is from the *Science Assessment Modules*, devised by Sue Harris and Steve Sizmur of the NFER. Once the invoice has been paid, it may be photocopied for use within the purchasing institution only. Published by The NFER-NELSON Publishing Company Limited, Darville House, 2 Oxford Road East, Windsor, Berkshire SL4 1DF, UK.

Code 4285 09 4

DSE in Brief

The German Foundation for International Development (DSE) is an institution which provides a forum for development policy dialogue and the initial and advanced training of specialists and executive personnel from developing and transitional countries. In addition, it supports German experts preparing themselves for their assignments in developing countries and maintains the Federal Republic of Germany's largest centre for documentation and information on development cooperation issues.

Conferences, meetings, seminars and training courses support projects which serve economic and social development, thus contributing to an effective, sustainable and wide-ranging development process.

The DSE cooperates with partners at home and abroad. A considerable number of the programmes take place in the developing countries, and the rest in Germany. Since 1960 the DSE has given advanced professional training to more than 120,000 decision-makers, specialists and executive personnel from over 150 countries. Through its dialogue and training events the DSE currently reaches more than 10,000 participants annually.

Founded in 1959, the DSE contributes to development cooperation on the basis of the guidelines of the German Federal Government's development policy. The institutional donor is the Federal Ministry for Economic Cooperation and Development (BMZ). Some of the DSE programmes are, however, financed by other donors (e.g. other Federal ministries, the Federal Länder, the European Union).

Also, the Federal Länder of Baden-Württemberg, Bavaria, Berlin, North Rhine-Westphalia and Saxony have made conference and training centres available. Since its establishment, the DSE has been jointly financed by the Federation and the Länder. This finds expression in the decentralized structure of the German Foundation with its specialized departments (Centres) and conference centres in a number of Federal Länder.