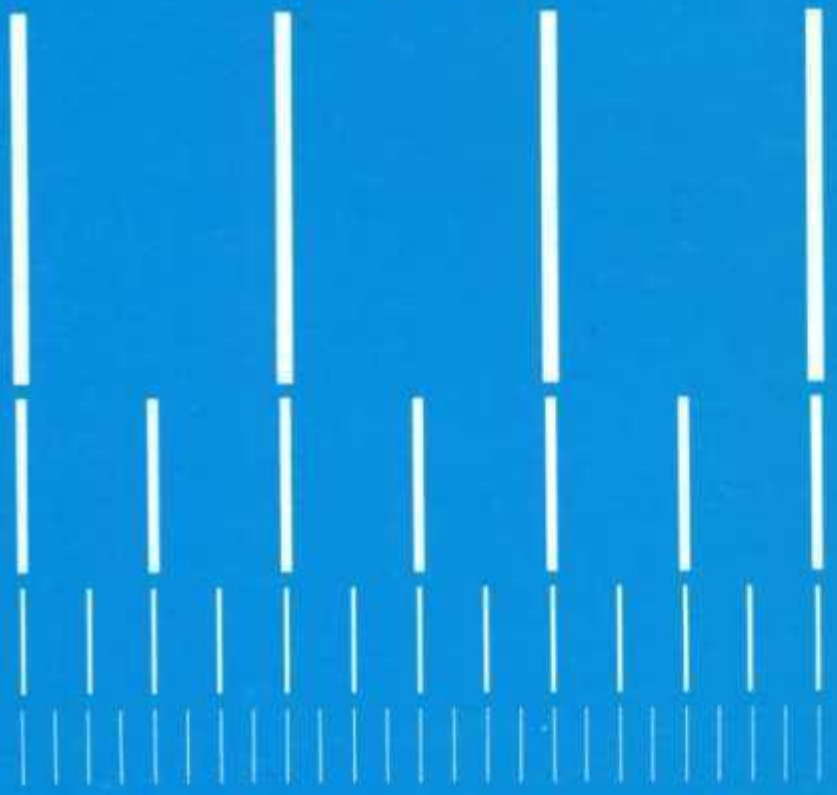
Fundamentals of Educational Planning

50

# National examinations: design, procedures and reporting

John P. Keeves



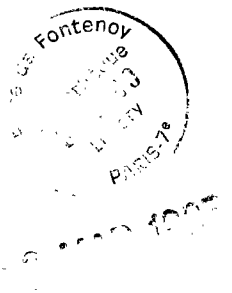
Paris 1994
UNESCO: International Institute for Educational Planning

Included in the series:*

2. The relation of educational plans to economic and social planning, *R. Poignant*
4 Planning and the educational administrator, *C.E. Beeby*
5. The social context of educational planning, *C.A. Anderson*
6. The costing of educational plans, *J. Vaizey, J.D. Chesswas*
7. The problems of rural education, *V.L. Griffiths*
8. Educational planning; the adviser's role, *A. Curle*
9. Demographic aspects of educational planning, *Ta Ngoc C.*
10. The analysis of educational costs and expenditure, *J. Hallak*
11. The professional identity of the educational planner, *A. Curle*
12. The conditions for success in educational planning, *G.C. Ruscoe*
13. Cost-benefit analysis in educational planning, *M. Woodhall*
18. Planning educational assistance for the second development decade, *H.M. Philips*
20. Realistic educational planning, *K.R. McKinnon*
21. Planning education in relation to rural development, *G.M. Coverdale*
22. Alternatives and decisions in educational planning, *J.D. Montgomery*
23. Planning the school curriculum, *A. Lewy*
24. Cost factors in planning educational technological systems, *D.T. Jamison*
25. The planner and lifelong education, *P. Furter*
26. Education and employment: a critical appraisal, *M. Carnoy*
27. Planning teacher demand and supply, *P. Williams*
28. Planning early childhood care and education in developing countries *A. Heron*
29. Communication media in education for low-income countries *E.G. McAnany, J.K. Mayo*
30. The planning of nonformal education, *D.R. Evans*
31. Education, training and the traditional sector, *J. Hallak, F. Caillods*
32. Higher education and employment: the IIEP experience in five less-developed countries *G. Psacharopoulos, B.C. Sanyal*
33. Educational planning as a social process, *T. Malan*
34. Higher education and social stratification: an international comparative study, *T. Husén*
35. A conceptual framework for the development of lifelong education in the USSR, *A. Vladislavlev*
36. Education in austerity: options for planners, *K. Lewin*
37. Educational planning in Asia, *R. Roy-Singh*
38. Education projects: elaboration, financing and management, *A. Magnen*
39. Increasing teacher effectiveness, L.W. Anderson
40. National and school-based curriculum development, *A. Lewy*
41. Planning human resources: methods, experiences and practices, *O. Bertrand*
42. Redefining basic education for Latin America: lessons to be learned from the Colombian Escuela Nueva, *E. Schiefelbein*
43. The management of distance learning systems, *G. Rumble*
44. Educational strategies for small island states, *D. Atchoarena*
45. Judging educational research based on experiments and surveys, *R.M. Wolf*
46 Law and educational planning, *I. Birch.*
47. Utilizing education and human resource sector analyses, *F. Kemmerer*
48. Cost analysis of educational inclusion of marginalized populations, *Mun C. Tsang.*
49. An efficiency-based management information system, *Walter W. McMahon.*

* Also published in French. Other titles to appear.

# National examinations: design, procedures and reporting




John P. Keeves

Paris 1994
UNESCO: International Institute for Educational Planning

The Swedish International Development Authority (SIDA) has provided financial assistance for the publication of this booklet.

Published in 1994 by the United Nations
Educational, Scientific and Cultural Organization
7 place de Fontenoy, 75700, Paris

Cover design by Bruno Pfäffli
ISBN 92-803-1154-9


IIEP/kof

# Fundamentals of educational planning

The booklets in this series are written primarily for two types of clientele: those engaged in educational planning and administration, in developing as well as developed countries; and others, less specialized, such as senior government officials and policy-makers who seek a more general understanding of educational planning and of how it is related to overall national development. They are intended to be of use either for private study or in formal training programmes.

Since this series was launched in 1967 practices and concepts of educational planning have undergone substantial change. Many of the assumptions which underlay earlier attempts to rationalize the process of educational development have been criticized or abandoned. Even if rigid mandatory centralized planning has now clearly proven to be inappropriate, this does not mean that all forms of planning have been dispensed with. On the contrary, the need for collecting data, evaluating the efficiency of existing programmes, undertaking a wide range of studies, exploring the future and fostering broad debate on these bases to guide educational policy and decision making has become even more acute than before.

The scope of educational planning has been broadened. In addition to the formal system of education, it is now applied to all other important educational efforts in nonformal settings. Attention to the growth and expansion of educational systems is being complemented and sometimes even replaced by a growing concern for the quality of the entire educational process and for the control of its results. Finally, planners and administrators have become more and more aware of the importance of implementation

strategies and of the role of different regulatory mechanisms in this respect: the choice of financing methods, the examination and certification procedures or various other regulation and incentive structures. The concern of planners is twofold: to reach a better understanding of the validity of education in its own empirically observed specific dimensions and to help in defining appropriate strategies for change.

The purpose of these booklets includes monitoring the evolution and change in educational policies and their effect upon educational planning requirements; highlighting current issues of educational planning and analyzing them in the context of their historical and societal setting; and disseminating methodologies of planning which can be applied in the context of both the developed and the developing countries.

In order to help the Institute identify the real up-to-date issues in educational planning and policy making in different parts of the world, an Editorial Board has been appointed, composed of two general editors and associate editors from different regions, all professionals of high repute in their own field. At the first meeting of this new Editorial Board in January 1990, its members identified key topics to be covered in the coming issues under the following headings:

1. Education and development
2. Equity considerations
3. Quality of education
4. Structure, administration and management of education
5. Curriculum
6. Cost and financing of education
7. Planning techniques and approaches
8. Information systems, monitoring and evaluation

Each heading is covered by one or two associate editors.

The series has been carefully planned but no attempt has been made to avoid differences or even contradictions in the views expressed by the authors. The Institute itself does not wish to impose any official doctrine. Thus, while the views are the responsibility of the authors and may not always be shared by UNESCO or the IIEP, they warrant attention in the international

forum of ideas. Indeed, one of the purposes of this series is to reflect a diversity of experience and opinions by giving different authors from a wide range of backgrounds and disciplines the opportunity of expressing their views on changing theories and practices in educational planning.

The present issue is devoted to the design, procedures and reporting of examinations. As planners and policy-makers are very well aware, examinations play a crucial role in educational systems. They provide the information on which selection for further education is based. They are also meant to certify pupils levels of competence for the use of future employers. Conscious of how important the passing of examinations is for a student's future life, parents, students and teachers often make enormous efforts, and sacrifice a considerable amount of time and/or money in order to maximize the chances of success. As a result, the content of public examinations often determines the teaching and learning that takes place in school, and educational systems are sometimes entirely dominated by them.

What is the role played by central examinations? What are their advantages and disadvantages? How could they be best conducted in an efficient and economic manner? How can the benefits be maximized and the losses that they induce be minimized? These are some of the questions, of direct interest to planners and policy-makers, that this booklet, prepared by Professor John P. Keeves of the Flinders University of South Australia, addresses.

Furthermore the Institute would like to thank Professor T. Neville Postlethwaite of the University of Hamburg, co-general editor and special editor of this series, for the active role he played in its preparation.

Jacques Hallak
Director, IIEP

*Composition of the Editorial Board*

# Preface

Nearly all countries of the world have school examinations. Some are set at the school level, some at the regional level and many at the national level. Some are used for certification purposes only, many are used for certification and selection purposes, and some are used for selection purposes only. Some systems of education rely on a central examination while others combine the results of a national examination with continuous assessment over a period of the pupil's life in school. Some examinations are of a criterion referenced kind but most are norm referenced. Where there is a national curriculum there is usually one central examination. However, with the decentralization of the curriculum in many countries there is a need to examine the pupils not only on the common core part of the curriculum but also on the local part of the curriculum and the local parts vary from one region to another. And the pupils' results on both parts have to be brought together on to a common scale.

The Editorial Board asked Professor John Keeves of the Flinders University of South Australia to prepare this booklet and, in particular, to concentrate on design, procedures and reporting. Professor Keeves has been a science teacher, President of the Australian Science Teachers Association, an educational researcher and the Director of the Australian Council for Educational Research. He also served for over a decade on the board of an examinations agency, the Australian Capital Territory Schools Accrediting Agency. This agency was developing a highly innovative programme for the accreditation of school curricula and for the moderation of school-based assessments. In these various

roles he has been associated with examinations. He has given much thought to the requirements for developing examinations, administering them, and marking them. More recently, he has been involved in ways of combining the examination of local and central curricula.

The IIEP is pleased that Professor Keeves accepted the challenge of preparing this booklet. It is the editors' view that readers will find the content to be of both practical and theoretical use.

T. N. Postlethwaite
Co-general editor

# Acknowledgements

In the development of this booklet I have drawn upon my career experiences as a student, classroom teacher, educational researcher, university academic, and consultant.

During this time I have had an opportunity to become involved with examinations agencies in Australia such as the Australian Capital Territory Schools Accrediting Agency. More recently, I participated in missions for the World Bank which were designed to foster wide debate on the planning and the use of national examinations in developing countries. In this latter area, I was greatly assisted by Dr Stephen P. Heyneman whose writings helped me to understand some of the major issues facing educators in developing countries.

I would also like to acknowledge those who taught me so much during courses on educational measurement at the University of Melbourne and in my work at the Australian Council for Educational Research, particularly the late Dr Wm. C. Radford and Dr S.S. Dunn. In addition, I am extremely grateful to Mr John Halsey of the Senior Secondary Assessment Board of South Australia and to Professor Dr T. Neville Postlethwaite of the University of Hamburg for their valuable comments on drafts of this booklet. Thanks must also go to Mrs Frances Anderson, who carefully and accurately typed the various drafts.

To those many people and agencies and to the *International Association for the Evaluation of Educational Achievement* (IEA) I owe the ideas present in this booklet.

John P. Keeves

# Contents


| | | |
|---|---|---|
| Preface | | 9 |
| Acknowledgments | | 11 |
| I. | Introduction | 15 |
| | Historical developments in examinations | 18 |
| | Changes in the design of examinations | 19 |
| | Persistent problems and the search for solutions | 21 |
| II. | National examination systems | 24 |
| | Review of policies and practices | 32 |
| | Effects of increased retention | 36 |
| | International baccalaureate | 38 |
| | Adult entry to higher education | 38 |
| | Conclusion | 39 |
| III. | Design, procedures and reporting of national examinations | 40 |
| | Role and function of school assessment | 41 |
| | Form of examination offered | 46 |
| | Design and mode of public examinations | 49 |
| | Procedures of curriculum planning and development | 53 |
| | Administrative problems in examinations agencies | 55 |
| | Equipment required by an examinations agency | 57 |
| | Procedures employed in the conduct of national examinations | 59 |

| | | |
|---|---|---|
| | Reporting of results from examinations | 65 |
| | Problems in the conduct of examinations | 68 |
| | Equity issues in the conduct of examinations | 70 |
| | Conclusion | 75 |
| IV. | Some technical issues | 76 |
| | Modelling levels of school achievement | 77 |
| | Allowance for differences in candidatures between subjects | 81 |
| | Moderation of school assessed marks between schools | 85 |
| | Item response theory scales | 89 |
| | Use of item response theory in banking | 93 |
| V. | National examinations – what is and what might be | 95 |
| | Problems of selection | 96 |
| | Problems of certification | 97 |
| | Problems of distortion | 98 |
| | Training of staff for an examinations agency | 99 |
| | Need for a sustained programme of research | 99 |
| | Conclusion | 100 |
| References | | 102 |

# I. Introduction

A national examination or a school leaving examination which assesses educational achievement has become one of the challenges that confronts young people at a particular stage in their lives in most countries of the world, provided they have had the opportunity to attend school. This opportunity is still denied to some children, while others leave school prematurely after only one or two years of attendance. For adolescent youth attending school a national examination serves two clear instrumental purposes. First, it provides information concerning levels of educational achievement on which selection is based to proceed to a further stage of education, or to enter an educational institution of a particular type such as a vocational or an academic school. Secondly, it provides information for certification of the levels of competence attained on particular educational outcomes. This certification influences entry into the workforce and the type of occupation in which employment might be sought. These two instrumental functions of national examinations, namely selection and certification are not only different in kind, but they also require different types of information from examinations. Selection requires the capacity to discriminate accurately near appropriate cut-off points, for example, a score of 482 out of 500, which is obtained by the top 2 per cent of candidates, and which might be required for selection into the prestigious faculty of medicine in an Australian university. Certification requires measuring attainment of a clearly specified standard of competence. It is, however, not clear that a single examination can always fulfil both functions. This conflict of

purposes results in tensions in the design, the procedures employed and methods of reporting the results of national examinations.

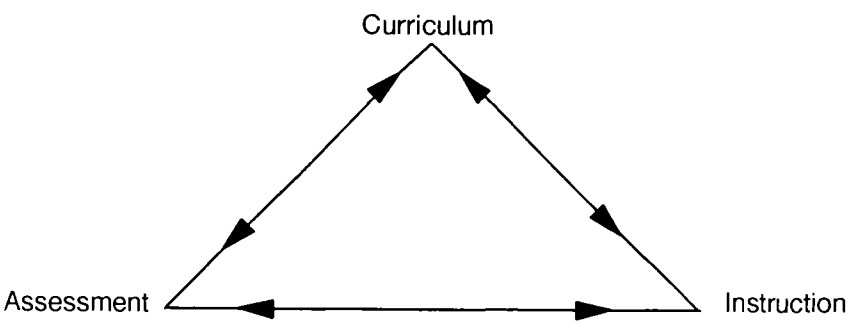
Advances have occurred in educational measurement during the years from 1960 to 1990 that have not only made clear what is involved in certification using a criterion-referenced approach, but have also sought, through using item response theory, to provide scales of measurement so that both functions of examinations could simultaneously be more effectively satisfied. It is necessary to recognize that among university academics, particularly in England, and those who maintain an elitist orientation towards higher education, these advances in educational measurement are challenged not on technical but largely on ideological grounds. It is the purpose of this booklet to address these issues, together with other aspects of the conduct of national examinations, and to place them in an historical and demographic context. In addition, this booklet seeks to draw attention to the educational benefits that might be made through the further development of the design of national examinations to meet changing societal needs and circumstances.

It is also necessary to recognize that national examinations have other functions besides those of selection and certification. The provision of education in all countries requires a significant component of public expenditure, commonly of the order of 5 per cent of GDP in OECD countries in 1988 (Bottani et al., 1992). It is thus not surprising that planners and policy-makers should be concerned not only with the curriculum of the schools but also with the standards of achievement of students under a particular curriculum. National examinations are thus seen as an instrument for control both of the curriculum and the instruction occurring within schools and as a mechanism for maintaining standards in education.

Some analysts such as Bernstein (1990) contend that graded individual performance and thus public examinations are used as instruments of social control in the education of youth emphasizing not only educational standards but also conformity to acceptable codes of behaviour. The contrast in the use of public examinations in Japan and the USA, both countries that hold a very high proportion of the age group at school to the twelfth grade level, may be seen to illustrate differences in social control between these

two countries. In both countries, however, there is currently reconsideration of the nature and functions of public examinations, in part, as a consequence of the reconceptualization of education as a lifelong or recurrent process.

In the redevelopment of education in countries of the Western World, many educators contend that the 'curriculum must come first'. This, however, fails to acknowledge the triangular and reciprocal relationships shown in *Figure 1* between the curriculum together with its aims and goals, instruction, and the assessment of achieved outcomes, as was so clearly advanced by Tyler (1949). Each of the three facets depends upon and interacts with the other two. Consequently, through the summative assessment of student learning by means of national examinations, control may be exerted on the curriculum, on instruction, and on codes of behaviour amongst youth.




*Source:* Tyler, R.W. 1949. *Principles of curriculum and instruction.* Chicago, University of Chicago Press.

Figure 1. The curriculum triangle

Acknowledgment of the relationships shown in *Figure 1* is not new. They have long been tacitly recognized, as can be seen from consideration of the development of examinations and the changing context within which such development has occurred. These changes are ongoing, and different countries are at different stages of development, not only in their conduct of national examinations, depending on contextual factors, but also in their approach to educational provision, and in the standards of education attained in their schools. Consequently, it is of profit to trace briefly the general worldwide development of national examinations. It should, nevertheless, be recognized that developments in the conduct of national examinations are not necessarily linear or uniform across countries. There is, however, sufficient in common across countries for the task of tracing the phases of development to be meaningful.

## *Historical developments in examinations*

While the conduct of written examinations can be traced back to the establishment of the Imperial Examination System in 606 AD in China (Lu Zhen, 1988) where such examinations were used for the selection of officials, it was only in the sixteenth century that the Jesuits introduced a system of competitive examinations into their schools and colleges across Europe (Madaus; Kellaghan, 1992). This subsequently led to the widespread use of written examinations in European universities alongside oral examinations as a way of raising the standards of education provided. Over time, the *abitur* examination emerged in Germany, the *baccalauréat* examination in France, and the *matriculation* examination in England. These examinations provided entry to the universities which were being established during the nineteenth century in order to recruit able young men into the professions and the public service on grounds of ability rather than family connection and wealth. The parallel developments across Europe ensured an emergence over time in the British, French and German colonies and in the USA of both universities and the use of written examinations for selection. Subsequently, the use of written examinations has become worldwide.

Simultaneously, with the gradual growth of elementary education, there was a concern that pupils should acquire the basic skills of reading, writing and arithmetic to prepare them for employment in commerce and industry. Teachers in some countries were considered accountable for their pupils' mastery of these skills, and examinations were held not only to provide certification for pupils, but also monetary rewards for teachers (Madaus; Kellaghan, 1992). Gradually universal primary education was introduced and expansion of the provision of secondary education, initially of an academic nature, also occurred. Examinations were then introduced to select pupils, who on completion of primary schooling would profit from secondary schooling. Where parents were unable to pay the necessary fees for secondary education, scholarships were provided for able pupils, who were selected on the basis of their performance in examinations at the end of primary schooling.

The differences that occurred between countries were not so much in the pattern of these developments, and in the functions of national and local examinations, but in the proportions of the age cohorts of children and adolescents for whom opportunities were available for education at successive stages of schooling either with the payment of fees, without payment of fees, or through scholarships and reduced fees. Examinations thus served the purposes of selection to balance demand and supply, and to identify able students whose parents could not afford the fees and living expenses required for further education. However, examinations also served purposes of certification for those who would terminate their education at a particular stage. In some countries, national or local public examinations were held each year during secondary schooling to form barriers on an age-grade educational ladder. In other countries, such examinations were held only at two or three-year stages and served as filters between stages of schooling.

## *Changes in the design of examinations*

This development of national or regional public examinations, while controlling standards of achievement, initially led to an emphasis on instruction in the knowledge and skills that could be

readily assessed in written examinations. This involved the learning of facts, drilling on computation, and the memorization of prepared answers to essay type questions. Such practices were generally seen to be contrary to the objectives of the curriculum, and to distort instruction. Moreover, as a school system gradually expanded, the work of conducting large-scale examinations, which involved a high proportion of the age cohort, tended to collapse under the weight of marking. This resulted in the elimination of some national examinations, and initially in the USA after World War I to the gradual use of objective tests that employed a multiple-choice format. Furthermore, in order to assist schools in assessing the performance of their students, and in the absence of comparative information from a national or regional examination, commercial publishing houses provided tests that had nation-wide norms. By means of these standardized tests which were administered under prescribed conditions, both schools and students were provided with standards of achievement against which to measure performance. The employment of such tests was gradually extended across the world, but with very different degrees of usage in different countries.

In the early years of their use, multiple-choice tests suffered from a tendency to construct test items in such a way that only knowledge of facts and principles was required to respond correctly to the items. In the 1940s, Tyler at the University of Chicago began experimenting with the construction of tests that assessed cognitive skills, which required understanding, application, analysis, synthesis and evaluation. The principles involved in the development of such tests were later laid down by Bloom (1956) in a taxonomy of educational objectives and extended to form a blueprint for test development and educational evaluation (Bloom; Hastings; Madaus, 1971). While these approaches had widespread acceptance, particularly in countries where the school system was expanding rapidly, both from marked increases in birth rates and from increased educational opportunities at all levels, they have tended to be rejected in continental Europe, where such pressures have not been as acute.

There has rightly been, throughout these developments, concern for the quality of learning taking place. This concern has involved

recognition that mere knowledge of facts and ability to spell and calculate accurately are insufficient and inadequate outcomes to be attained as a result of many years of schooling. Modern society demands more.

## *Persistent problems and the search for solutions*

The 150 years from the middle of the nineteenth century to the closing decade of the twentieth century has seen remarkable growth in education. In successive stages, there has been the initial expansion of the first phase of education, primary schooling, followed by the expansion of the second phase of education or secondary schooling. In some countries, these developments have involved a separation of academic schools from vocational schools, and in other countries the establishment of comprehensive schools. As a result of these developments, there are in the mid-1990s some countries that enrol over 90 per cent of the age group at the twelfth grade level. Alongside this growth has been the expansion of university education, until a stage has been reached in the mid-1990s, where in some countries over 40 per cent of an age group enter universities or higher education for tertiary studies. In addition, a further proportion, up to 36 per cent of an age group in certain countries, undertake further training in technical colleges (Bottani *et al.*, 1992). Furthermore, a strong view is also emerging of education as a lifelong or recurrent process, and this demands reconsideration of the manner and nature of all national examinations.

Different countries have progressed to differing degrees along these paths. There is clearly a place for selection at a national-level where opportunities are limited. Nevertheless, it must be asked whether there is the same need for national examinations with an emphasis on the selection function in a very open educational system. There is possibly a greater need in such countries for a strong emphasis on the certification function in national examinations to indicate whether a student is capable of and adequately prepared to advance to the next stage, or is competent to undertake the work involved in a trade or other occupation. Associated with these functions of national examinations are the issues of whether

or not they operate fairly and with equitable fairness with respect to socioeconomic background, sex, or ethnic group to which an individual student may belong. These problems demand answers, but not necessarily the same answers for each country, nor the same answer for all times within a specific country. It is the purpose of this booklet to address these questions and to suggest ways in which answers might be sought within a particular context.

The context for a particular country is not only influenced by educational or economic development, but also by the dominant political ideology of that country. The government of a particular country with a clearly identifiable social agenda may demand changes in the public examination system that operates within that country, perhaps to fulfil purposes of social control or alternatively to ensure greater equity for all subgroups in the society. Such changes are not necessarily those that might be argued by educators to be the most educationally desirable. However, an operational consensus must be achieved at any given time or the educational system would collapse, although it is ultimately the government of the day that must determine educational policy and as a consequence educational practice.

Educational objectives and aims also change. In the mid-to-late-1970s, in certain countries, the role of the secondary school was being challenged (Husén, 1979). However, in the following decade these doubts were removed and both policy-makers and youth by the mid-1990s were no longer equivocal about the value of education for either personal development or for national advance, particularly in science-related fields. The educational problems being raised in many countries in the 1990s are concerned with whether national examinations can be employed to promote both personal and national development and if so, how are such goals best achieved? Inter-relations were argued above to exist between the curriculum and its objectives, the instruction provided and the assessment of outcomes. There is, nevertheless, the issue as to how examinations can best be conducted in order to augment both the curriculum objectives and the instructional programme through the appropriate assessment of student performance

With huge numbers of students moving through educational courses in some countries, the question must be raised as to how examinations might best be conducted both efficiently and economically to maximize the benefits to be gained and to minimize the losses. Answers must be sought through the procedures employed in the conduct of public examinations. Here, previous experience is likely to be of assistance in the identification of procedures that are advantageous. Nevertheless, the use of certain procedures requires training of examinations agency staff and teachers in schools to a satisfactory level of competence. Without such training the use of such procedures is likely to be ineffective.

In *Chapter II* the range of procedures employed across different countries is briefly considered. In addition, suggestions are made as to why these different arrangements have developed. In *Chapter III*, this booklet considers both promising procedures and the necessary training to employ these procedures. Some of the issues raised require technical knowledge and understanding, and if pursued are likely to demand further and on-going training for those who employ these procedures. In *Chapter IV*, these technical issues are briefly addressed, and references are provided from which further information might be obtained. The technical problems considered include those of:

- allowance for differences in retention rates;
- the equating of school assessments;
- the rescaling of scores for differences between candidatures;
- the development of item response theory scales of achievement; and
- use of item banks in public examinations.

In the concluding chapter, some discussion is provided on the need for maintaining a programme of research into the intended and unintended consequences of a particular national examination. Without such research, an examination system is poorly placed to respond to changing needs and circumstances.

# II. National examination systems

There is no common approach across countries to the conduct of national examinations. Different countries at different stages of development and in different historical contexts employ different approaches to the selection and certification of young people during and on completion of their schooling. Nevertheless, from the evidence that was assembled in the Second IEA Science Study (see Rosier; Keeves, 1991; Postlethwaite; Wiley, 1992; Keeves, 1992) and in studies conducted by The World Bank (Heyneman; Fägerlind, 1988; Za'rour, 1987) and by comparative educators (Eckstein; Noah, 1992; 1993) it is possible to detect certain patterns, relationships and trends in the manner in which countries design and conduct public examinations at a national- and sub-national-level. In this chapter, the evidence available is briefly examined in an attempt to understand the developments and changes that have occurred over both time and location in national examination systems.

Some selection of the countries considered is necessary. There are simply too many countries and it is too difficult to assemble accurate information for all countries that might be examined. Consequently, some choices must be made from among the developed and developing countries. The IEA National Science Case Studies, which were prepared in the mid-1980s and were based on the years 1983-1984, provide a valuable source of information (Rosier; Keeves, 1991). However, even here, the data available are incomplete, in spite of extended correspondence and sustained efforts. The countries involved in this investigation include many highly developed industrialized nations, such as Japan and the USA,

as well as developing countries from Africa, Asia, and Oceania. Nevertheless, no Latin American countries, no Francophone countries and no Arab States are included. While this is to be regretted, it does not necessarily prevent the building of an understanding of the factors that influence the developments and changes that occur over time in national examinations systems. More serious perhaps is the fact that the evidence assembled is cross-sectional in nature and not longitudinal, and that changes occurring over time must be inferred and not examined directly. The factor that would seem to have a dominant influence on the changes taking place over time is the growth in participation rates within each country at successive stages of schooling which has occurred during the twentieth century. Thus, for the purposes of comparison, it is necessary to have a variety of countries that differ widely in their participation rates at the different stages of schooling and to examine whether they deal with the problems of selection, certification and system evaluation in different ways. It should of course be noted that, in situations where all or nearly all students in an age group continue with education beyond a particular stage, there is little need for selection, unless such selection is required to allocate students to different types of schools. Issues of certification, system accountability and curriculum evaluation remain.

In *Table 1* information relating to the conduct of public examinations is recorded for 24 school systems which relates to the 1983-1984 school year. This information is presented with full recognition that significant changes may have taken place in particular countries since that time. For further details, the reports of the Second IEA Science Study (Rosier; Keeves, 1991; Postlethwaite; Wiley, 1992; Keeves, 1992) should be consulted. In particular, the case studies of science education, and the introductory discussion to these case studies recorded by Rosier; Keeves (1991, Chapter 2) contain much of the detail and the necessary qualifications to the summary data presented in *Table 1*.

*Table 1* gives the participation rates for three specific age and grade levels of schooling, namely the middle primary grades, the middle secondary grades and the terminal secondary school level. Countries differ in the age at which schooling commences.

Table 1. Information relating to educational participation at successive stages of schooling and the conduct of national examinations

| Stages | Mld-primary | | | Mld-secondary | | | Terminal secondary | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Country (IEA) | % | Grade | Age | % | Grade | Age | % | Grade | Age | School type |
| Australia | 99 | 5 | 10:6 | 98 | 9 | 14:5 | 39 | 12 | 17:3 | G |
| Canada (English) | 99 | 5 | 11:1 | 99 | 9 | 15:0 | 64 | 12/13 | 18:3 | G |
| Canada (French) | 99 | 5 | 11:1 | 94 | 9 | 15:3 | 67 | 11/12 | 17:2 | A/V |
| China | na | na | na | 37 | 9 | 15:8 | na | na | na | A |
| England | 99 | 5 | 10:3 | 98 | 9 | 14:2 | 20 | 13 | 18:0 | A |
| Finland | 99 | 4 | 10:10 | 99 | 8 | 14:10 | 63 | 12 | 18:6 | A/V |
| Ghana | - | - | - | 43 | 9 | 16:1 | 1.2 | 13 | 18:8 | A |
| Hong Kong | 99 | 4 | 10:5 | 99 | 8 | 14:7 | 20 | 13 | 19:2 | A |
| Hungary | 99 | 4 | 10:3 | 92 | 8 | 14:3 | 40 | 12 | 18:0 | A/V |
| Israel | 99 | 5 | 10:9 | 99 | 9 | 14:9 | 75 | 12 | 17:6 | G |
| Italy | 99 | 5 | 10:9 | 99 | 8 | 13:11 | 34 | 13 | 19:0 | A |
| Japan | 99 | 5 | 10:7 | 99 | 9 | 14:7 | 89 | 12 | 18:2 | A/V |
| Korea | 99 | 5 | 11:2 | 99 | 9 | 15:0 | 83 | 12 | 17:9 | A/V |
| The Netherlands | na | na | na | 99 | 9 | 15:6 | na | na | na | A/V/G |
| Nigeria | 92 | 6 | 12:1 | na | 10 | 16:2 | na | na | na | A |
| Norway | 99 | 4 | 10:11 | 99 | 9 | 15:10 | 40 | 12 | 18:9 | A |
| Papua New Guinea | - | - | - | 11 | 10 | 17:1 | 1.1 | 12 | 18:7 | A |
| Philippines | 97 | 5 | 11:1 | 60 | 9 | 16:1 | - | - | - | 6 |
| Poland | 99 | 4 | 10:11 | 91 | 8 | 15:0 | 28 | 12 | 18:6 | A/V |
| Singapore | 99 | 5 | 10:10 | 91 | 9 | 15:3 | 22 | 12/13 | 18:1 | A |
| Sweden | 99 | 4 | 10:10 | 99 | 8 | 14:10 | 28 | 12 | 19:0 | A |
| Tailand | - | - | - | 32 | 9 | 15:4 | 29 | 12 | 18:3 | A/V |
| USA | 99 | 5 | 11:3 | 99 | 9 | 15:3 | 80 | 12 | 17:7 | G |
| Zimbabwe | na | na | na | 30 | 9 | 16:1 | na | na | na | NA |

Table 1. (continued)

| Stages | Grade at end of stage | | | Exam at end of stage | | |
|---|---|---|---|---|---|---|
| Country (IEA) | Primary | Secondary | Upper secondary | Primary | Secondary | Upper secondary |
| Australia | 6/7 | - | 12 | - | - | 12s/S |
| Canada (English) | 6 | 8/9 | 12/13 | - | - | 12s/S |
| Canada (French) | 6 | 9 | 11/12 | - | - | 12s/S |
| China | 5/6 | 8/9 | 11/12 | t | t | 11/12 |
| England | 6 | - | 13 | - | 10/11 | 13 |
| Finland | - | 9 | 12 | - | - | 12 |
| Ghana | 6 | - | 13 | na | na | na |
| Hong Kong | 6 | 9 | 13 | S | S | 11,12,13 |
| Hungary | - | 8 | 12 | - | - | S |
| Israel | 6 | 9 | 12 | - | - | 12 |
| Italy | 5 | 8 | 13 | - | - | S |
| Japan | 6 | 9 | 12 | - | t | 12,U |
| Korea | 6 | 9 | 12 | - | t | S,U,12 |
| The Netherlands | 6 | - | 12 | - | - | 10/11/12 |
| Nigeria | 6 | 9 | 12 | 6 | 9 | 12 |
| Norway | 6 | 9 | 12 | - | S | 12s |
| Papua New Guinea | 6 | - | 12 | 6 | 10 | 12 |
| Philippines | 6 | - | 10 | - | - | 10 |
| Poland | - | 8 | 12 | - | t | 12/U |
| Singapore | 6 | 10 | 13 | 3,6 | 10 | 12,13 |
| Sweden | - | 9 | 12 | - | 9s | 12s |
| Tailand | 6 | 9 | 12 | - | - | U |
| USA | 6 | 8 | 12 | | C | C,E |
| Zimbabwe | - | - | 13 | na | na | na |

Table 1. (continued)

| Stages | Tertiary education | | Terminal secondary | |
|---|---|---|---|---|
| Country (IEA) | % higher/secondary | % non-university | % male | ù female |
| Australia | 35 | na | 36 | 42 |
| Canada (English) | na | na | 63 | 65 |
| Canada (French) | na | na | na | na |
| China | 2 | na | na | na |
| England | 15 | 6 | 20 | 20 |
| Finland | 28 | 24 | 31[a] | 49[a] |
| Ghana | na | na | na | na |
| Hong Kong | na | na | na | na |
| Hungary | na | na | 13[a] | 23[a] |
| Israel | na | na | 65 | 84 |
| Italy | 27 | 1 | na | na |
| Japan | 25 | 27 | 87 | 92 |
| Korea | na | na | 79 | 88 |
| The Netherlands | 12 | 18 | na | na |
| Nigeria | na | na | na | na |
| Norway | 16 | na | na | na |
| Papua New Guinea | na | na | 2 | 1 |
| Philippines | na | na | na | na |
| Poland | na | na | na | na |
| Singapore | na | na | 19 | 26 |
| Sweden | 14 | 36 | 28 | 28 |
| Thailand | na | na | 12[a] | 16[a] |
| USA | 48 | 22 | 75 | 85 |
| Zimbabwe | na | na | na | na |

*Notes*: The information recorded in *Table 1* has been compiled, in the main, from data collected in the Second IEA Science Study which was conducted in 1983-84. Thus the information recorded relates to those years, and is presented with full recognition of the significant changes that have taken place in many countries since that time. For further details the reports of the Second IEA Science Study (Rosier and Keeves, 1991; Postlethwaite and Wiley, 1992; Keeves, 1992) should be consulted. In particular, the case studies of science education and the introductory discussion to these case studies recorded by Rosier and Keeves (1991) contain much of the detail and the necessary qualifications which relate to the material presented.

| | | |
|---|---|---|
| na | = | no data available. |
| age | = | recorded in years and months. |
| / | = | indicates an or relationship. |
| a | = | academic schools only. |
| A | = | academic schools. |
| V | = | vocational schools. |
| G | = | schools providing a general education. |
| , | = | to indicate an and relationship. |
| S | = | only school assessments are employed. |
| s | = | school assessments are combined with examination marks. |
| U | = | universities conduct their own examinations for selection. |
| E | = | commercial agencies conduct examinations for selection. |
| C | = | minimum competency tests are administered at a state level for certification purposes. |
| t | = | schools or district boards conduct selection examinations for entry to the next stage of education. |

Information recorded in *Table 1* on tertiary entrance, which involves the ratios of first-time entrants into full-time public and private tertiary education as a proportion of the 1988 age cohort, were obtained from Bottani *et al.* (1992). Data are given for two sub-groups namely, the percentage in higher (university) education, and the percentage in non-university tertiary education.

Age of entry ranges from five years in England, to seven years in the Nordic countries, and with a range from six to eight years in Nigeria, Ghana and Zimbabwe. Likewise countries differ in the number of years required to reach the terminal grade, which was, in 1983-1984, Grade 10 in the Philippines, Grade 11 in the province of Quebec, Grade 12 in a majority of countries, and Grade 13 in England and those countries that followed an English model, such as Ghana, Hong Kong, Singapore, and Zimbabwe. For those countries that provided data, there are over 90 per cent of the age group in school at the mid-primary stage. However, opportunities for secondary education are limited in many countries. The highly industrialized countries have an estimated 99 per cent of the age group in school at the middle secondary school level. However, the proportion of the age group participating in that level of schooling falls to 11 per cent in Papua New Guinea, 30 per cent in Zimbabwe, 32 per cent in Thailand, 37 per cent in China, 43 per cent in Ghana, and 60 per cent in the Philippines. It is of interest to consider both how those who attend are selected in these six countries, and whether those who have dropped out have received a certificate to indicate the level of schooling attained.

By the terminal secondary school stage there has been a further decline in participation rates, together with an earlier choice in many countries to continue with schooling in either an academic or a vocational programme. Where places are limited both in upper secondary schools or in a particular type of secondary school, some selection is required prior to entry to the further stage of secondary education. Participation rates at the terminal year level ranged widely from over 80 per cent of the age group in Japan, Korea, and the USA to approximately 1 per cent in Ghana and Papua New Guinea. To some extent the selection that took place was self-selection by the students or a process of 'cooling off' operated by the schools. Nevertheless, in approximately half the countries for which there is information some type of selection testing or a national examination operated at the end of the middle secondary school stage in order to select students for upper secondary education. In five of the countries considered, a similar selection process operated, in 1983-1984, at the completion of primary schooling prior to entry into secondary education.

The data are very incomplete for the proportion of an age group that can be considered to be first-time entrants into full-time tertiary education whether of a university or non-university type and relate to 1988 (see Bottani *et al.*, 1992). The participation rates in full-time university or higher education range from 48 per cent in the USA, an estimated 35 per cent in Australia, 25 per cent in Japan, 15 per cent in England, to 14 per cent in Sweden. However, in full-time non-university education, the first-time entrants range from 36 per cent in Sweden, 27 per cent in Japan, 22 per cent in the USA, to 16 per cent in England, with no estimates available for Australia. The mechanisms of selection into these two types of education at a stage beyond secondary schooling warrants consideration, even though for so many countries of interest this basic information and the proportions of the age group involved would not appear to be available.

On completion of the terminal stage of secondary schooling, in all countries for which information was available, there was some type of national examination or its equivalent in place in 1983-1984. These programmes of assessment serve the purposes of either selection for entry into training or higher education or for certification for entry into the labour force. The procedures employed, in addition to an external examination conducted by a national agency (indicated by a Grade number) include an examination conducted by a commercial agency (E), examinations conducted by particular universities (U), school assessments and school examinations only (S), school assessments scaled and equated by national examination marks (s), and minimum competency tests both for certification and high school graduation (C). It is evident that a wide variety of approaches is employed.

It should be noted that on the one hand the developing countries for which evidence is available are the countries that maintain an extensive examination system at the conclusion of each of the major stages of schooling, for example, China, Nigeria, Papua New Guinea. In a similar way – Singapore but not Korea, both countries that have developed their industrial base since 1950 – conducts examinations at the end of each stage of schooling. It is of interest to consider why Korea should be different. The answer

may lie in the high proportion of the age group that remains to the twelfth grade in Korea.

On the other hand the USA, a country that holds over 80 per cent of the age group to the terminal stage of schooling in a majority of states, separates the selection function of the terminal examinations from the certification function and maintains a two stage process, with high school graduation largely, although not completely, separated from university entry. Likewise in Japan, a country with high retention rates to the terminal grade, the national and local public universities, and some of the large number of private universities, maintain a two stage examination programme. Common achievement examinations are administered at the first stage, by the National Center for University Entrance Examinations (NCUEE). Subsequently, many universities offer their own selection examinations at a second stage. The first stage examination aims to evaluate applicants' attainment in basic and general studies in upper secondary schools, largely a certification function. The second stage examination assesses the abilities and aptitudes of candidates and has a selection function for entry into a particular institution (Hidano, 1988). Certificates of graduation from the upper secondary school are also issued by individual schools on the basis of completed school courses to the satisfaction of the teacher in charge of the course.

Other countries with a relatively high proportion of the age group remaining at school, for example, Australia, Canada (English), Norway and Sweden, depend heavily on school assessments. These school assessments may involve school-based internal examinations as well as continuous assessment programmes and graded assignments throughout the final years of schooling. They are combined to provide information on which to base both selection and certification at the completion of the final stage of schooling.

## *Review of policies and practices*

The discussion and evidence presented above on the conduct of public examinations consider the arrangements provided for selection at successive stages of education in the school systems

that took part in the Second IEA Science Study. It would seem possible to draw several important general conclusions about the conduct of public examinations in relation to the educational context of a country from this evidence. Nevertheless, there are countries not on the list of 24, which would be of value to discuss. In France, for example, university entry is open for all who pass the national examination at the end of an intensive upper secondary school programme. In Brazil, the extensive private educational system both at school and university level presents particular problems. However, the unique conditions in each of these countries, while of considerable interest, may not help to provide greater understanding of the factors that influence the design, procedures and reporting employed in national examinations.

1. The holding of a high proportion of an age cohort through successive stages of schooling is both a consequence of economic and technological development and serves to facilitate further economic development. Thus, in countries with high retention rates the school system has responded to changing labour market requirements by retaining a high proportion of the age cohort at school. However, this necessitates that the school system also provides appropriate education and training in changing circumstances. There were, in 1984, interesting exceptions where, for example, Korea, which has undergone rapid industrial growth in the latter half of the twentieth century, retained a very high proportion of the age group at school to Grade 12 (83 per cent), while in England only a relatively small proportion of the age group (20 per cent) was retained to the final secondary stage.
2. Where a high percentage of an age cohort proceeds to successive stages of education, the need for selection is reduced or eliminated. This reduction of need may lead to a consequent reduction of emphasis on public examinations. However, there is need to consider issues of certification that would maximize post-graduate options and educational pathways to be followed later in life.
3 Where the demand for places in educational programmes at a higher level exceeds the supply of places, the need exists for

some type of selection procedures to identify and allocate students to programmes. Selection examinations are also used to resolve the demands for entry into particular courses, where the number of places is limited.

4. Likewise, where the demand for places is limited in one type of institution, for example, academic schools in contrast to vocational schools, selection into the different types of schooling is required. In some countries, regional examinations, as are conducted in Japan, are employed for this purpose. However, results on these examinations may also serve the purposes of selection into more prestigious schools in those countries where competition is intense at a later stage of education.
5. At all stages of schooling, public examinations generally serve purposes of both selection and certification. Where a high percentage of an age group is retained to the terminal secondary school stage as well as in higher education, the demands of certification for entry into the labour market as well as those of selection into training programmes or more prestigious institutions of higher education and more prestigious courses, would appear incompatible. As a consequence, in countries where such circumstances apply, the examinations conducted at the terminal secondary school stage may operate at two levels. At one level, the examinations serve purposes primarily of certification; at a second and higher level, the examinations clearly serve purposes of selection. In some countries, such as Australia, where rapid growth in upper secondary enrolments has taken place in recent years, this problem is not adequately resolved: only one assessment operation is required to serve both purposes.
6. The size of an educational system presents problems in the conduct of public examinations, both as a consequence of the associated diversity of schools, and the magnitude of the operation. The responsibility for the conduct of examinations is thus delegated in some countries to sub-units, such as state or regional authorities. However, where there is considerable mobility across state or regional boundaries within a country, then some nation-wide body is required. In the USA, the

boards conducting examinations are national commercial or non-profit bodies. In China, a national organization operates, but under considerable strain, and has sought ways of reducing that strain (Heyneman; Fägerlind, 1988).

7. Another consequence of the magnitude of the public examination enterprise in certain countries, is the employment of objective tests that are machine marked. This may, however, lead to a reduction of emphasis on the writing of coherent and extended prose, with a lowering of the students' skill on such tasks, as has been said to occur in the USA (Linn *et al*, 1991).
8. In order to assess performance with respect to skills, other than those measured by objective tests, there has been gradual inclusion of a school assessment component into the marks awarded to students. School assessment procedures permit, in the awarding of marks, consideration of performance in the science laboratory, the writing of extended reports and assignments in the social sciences and the humanities, and the creation of an original piece of work in the fields of art and music. Moreover, they permit the greater sampling of performance, beyond that possible in an examination of only a few hours. Some countries, such as Sweden and certain regions of Australia and Canada, have all but eliminated public examinations in preference for a reliance on school assessment of student performance. There is, nevertheless, the need to maintain comparability across schools, and in these countries procedures of moderation involving the use of scholastic aptitude tests or reference achievement tests are commonly employed for this purpose.
9. Countries, where an educational system is being developed and rapidly expanded, have concern for the maintenance of the standards of education provided. Singapore, China and Papua New Guinea are countries where such concerns would appear to exist. In these countries, the conduct of public examinations at several levels is used deliberately to influence educational practice in schools and to raise the level of education.
10. One of the mechanisms for increasing participation in education at successive stages of schooling, as in Australia and in Sweden, has been the removal of barriers in the form of public

examinations, and practices of grade repetition. Thus, the gradual elimination of public examinations and grade repeating in countries where such policies and practices were formerly well-established, has served to increase participation in education. However, it is also necessary for the curriculum to be redesigned and for instructional procedures to change.

It is evident that public examinations have a more dominant role in developing and newly industrialized countries than in the highly developed countries at the intermediate stages of schooling, if not at the terminal stage. However, Korea and Thailand on the one hand would appear to be exceptions to this general rule. Moreover, Korea has achieved a high level of educational participation to the end of 12 years of schooling, while Thailand has not. On the other hand, Japan and England would appear to be exceptions among the more highly developed countries. Furthermore, Japan has achieved a high level of participation, at the terminal level of schooling, while England has not.

The 10 generalizations listed above are beset with exceptions, but as trends or tendencies they would appear valid. This booklet is concerned with national examinations, and their design, procedures and reporting. Thus, it may seem to be more relevant to the developing countries than to the more highly industrialized ones. However, even this general comment requires some qualification, and further consideration must be given to the effects of increased retention rates and to the effects of the introduction of programmes of adult, recurrent and lifelong education.

## *Effects of increased retention*

One of the consequences of increased retention rates is that there is a greater pool of more highly educated personnel from which a country may draw in its quest for skilled professional, technological, commercial and industrial workers to support economic growth. One of the concerns of policy-makers must inevitably be whether the expansion of an educational system leads to a decline in standards, particularly among the more able who must lead and promote further development. This is a difficult

question to answer, and only the studies carried out by the *International Association for the Evaluation of Educational Achievement* (IEA) would appear to have provided evidence which permits this issue to be addressed. While these studies do not involve national examinations directly, they use sample surveys which assess educational achievement to investigate relationships across countries, in the absence of appropriate common examinations.

IEA studies have repeatedly recorded a negative relationship across countries between mean levels of achievement at the terminal secondary school level in mathematics (Husén, 1967) and science (Comber; Keeves, 1973; Postlethwaite; Wiley, 1992; Keeves, 1992) and participation rates in the study of the particular subject. Thus, it would appear that as retention and participation rates increase across countries, and by inference within a particular country, the average level of achievement of the students tested might be expected to decline. A model was developed by McIntosh (1959) for the effects of selection or retention on standards of achievement in large populations, and is considered in more detail in *Chapter IV*.

It is also important to note that IEA studies have shown that the performance of the more able students does not necessarily decline the higher the proportion of the age group retained (Postlethwaite, 1967; Husén, 1967; Comber; Keeves, 1973; Keeves, 1992; Postlethwaite; Wiley, 1992). Thus, the average level of achievement may fall but the achievement of a fixed proportion of the age group should not be expected to fall when a higher proportion of the age group is retained at school. During the past century standards in higher education may well have risen substantially as the number of years of schooling provided has gradually increased. In countries where further years of schooling are added, as is occurring in Canada (Quebec Province) and China, and might be anticipated to occur in the Philippines, even though the proportion of the age group retained is increased, the standards might be expected to rise.

Only in the USA have attempts been made to examine changes over time in standards of achievement on examinations that are widely used for selection into higher education (Donlon, 1984). In these studies the confounding effects of fluctuations in participation

rates are evident. The increased yield of a school system in a particular subject which results from increased retention and participation rates, might be expected to have beneficial effects on national scientific, technological and economic development (Husén; Keeves, 1991).

## *International baccalaureate*

In the preceding section, consideration has been given to the variety of examination provisions made in different countries and to the selection of students for entry into higher education within a particular country. However, many universities in highly developed countries make provision for entry of up to 10 per cent in their annual intake of students from other countries and the Erasmus Scheme operating in European countries supports the exchange of students between countries. It is generally argued that a mix of students from different cultures contributes significantly to the quality of life and thought within a university. In addition, where such students pay fees, the financial gains to an institution are also of considerable benefit. To facilitate the movement of students across countries, the *International Baccalaureate* (IB) has been established with, in 1983, affiliated schools that teach IB courses and institutions that accept IB qualifications in 51 countries. Examinations conducted by IB are of high standard and demand a rich and scholarly, but largely traditional, school programme (Halls, 1985). However, IB courses are designed to serve the needs of the more able students.

## *Adult entry to higher education*

The emerging view of lifelong or recurrent adult education, that has flowed from the UNESCO publication *Learning to be. The world of education today and tomorrow.* (Faure, 1972), has led in many countries to provision for mature-aged entrants into higher education. Commonly a period of employment followed by participation in a preparatory programme are the requirements for adult entry to higher education. While there is often evidence of strikingly successful performance by mature-aged entrants, such

success is not common in those fields that involve the study of mathematics and the physical sciences. Adult entry does not, in general, require that applicants should take a national examination, although some enter higher education on the basis of their performance in such an examination.

## *Conclusion*

The great variety in examination programmes across countries reflects the diversity in the conditions and context of educational provision, the variation in student ages at different grade levels, as well as variation in participation rates at key stages of schooling in different countries. It is recognized that in addition to serving the purposes of selection and certification for students, national examinations, though prescribing the curriculum of the schools, have the possibility of advancing or retarding the quality of education in the schools of a particular country. Moreover, through recognition of particular outcomes, they have the possibility of either raising or lowering the quality of teaching provided in schools. In the chapter that follows, consideration is given to the more practical problems involved in the design, procedures and reporting of national examinations that could help to promote higher quality public education, particularly in those countries where greater use is made of such examinations at more than one grade level, commonly the developing countries.

# III. Design, procedures and reporting of national examinations

National examinations place pressures on schools, teachers and individual students. Sometimes these examinations create individual anxieties and tensions, and these may, under certain circumstances, adversely affect the results of individual students as well as the way schools and their teachers undertake their work. Nevertheless, such tensions may be employed to stimulate the work of schools, through providing direction, incentives, and rewards in the form of increased student performance. As a consequence, national examinations should be designed in order to profit from these tensions and to use them to raise the quality of education that every school within the system offers.

There are many aspects of the conduct of public examinations where choices have to be made that will, under different circumstances, have differing effects on the quality of education provided in the schools. Decisions made are commonly quasi-political in nature and are influenced by tradition and pressure groups, such as are formed by an older generation of academics who themselves grew up under a particular system and are resistant to change, or by militant teachers who received their university training during the years of student radicalism of the late 1960s and early 1970s. Those aspects that would seem to warrant consideration in a discussion of testing dilemmas and that might influence the quality of education provided in response to a national examination include the:

(i) Role and function of school assessments;
(ii) form of examination offered;
(iii) design and mode of public examinations;

(iv) procedures of curriculum planning and development;
(v) administrative problems in examinations agencies;
(vi) equipment required by an examinations agency;
(vii) procedures employed in the conduct of national examinations;
(viii) reporting of results from examinations;
(ix) some problems in the conduct of examinations; and
(x) equity issues in the conduct of examinations.

Each of these 10 aspects is considered in turn on the pages that follow. In addition, the linkages with other sectors of the educational system must be considered in relation to these 10 aspects of the conduct of national examinations.

A discussion of the range of options available exposes the dilemmas confronting policy-makers and practitioners who are faced with a need to make decisions in a particular context, and commonly without recourse to research. The variety of practices across the 24 school systems considered in the previous chapter serves to indicate the difficulty of making such decisions. Nevertheless, certain decisions may be shown to have beneficial effects on educational practice and other decisions may be shown to have effects that are detrimental.

Throughout the discussion that follows there is need to consider what an ideal or effective examinations centre might look like. Both the efficiency with which an examinations centre carries out its work and the extent to which its activities are educationally beneficial are influenced by:

- how it is organized;
- its size in terms of professional and support staff;
- its relations with other sectors of the educational system;
- the financial support it receives and the equipment it uses; and
- how it goes about the design and development of the examinations it conducts.

## *Role and function of school assessments*

Teachers generally advocate the use of school assessment-procedures in preference to external examinations. In contrast,

institutions of higher education commonly support the use of external examinations. As a teaching service gains in professionalism and strength within a country, it demands the right to employ school assessments in opposition to university pressures to the contrary. It is argued that school assessments reward industry, motivation and the will to succeed. Moreover, it is claimed that school assessments permit the marking of aspects of performance that cannot be measured in pencil and paper examinations of limited duration. Thus, curriculum objectives that would otherwise be ignored if school assessments were not used are said to be able to be rewarded and fostered. However, the problems of marking student performance with consistency and in a meaningful way, sometimes leads to such outcomes being assessed for certification but not assessed for selection purposes.

*Models of school assessment*

The tasks or questions assessed within schools or in public examinations may be prepared in different assessment modes. The commonly used modes of assessment are:

- objective test items, largely of the multiple-choice type;
- short answer questions, where either a calculation must be carried out or a brief response constructed;
- essay type questions, where the student must write an extended response; and
- a performance task or a problem solving exercise must be carried out.

There is a growing demand that school assessment should place increased emphasis on the third and fourth modes, while public examinations are increasingly employing the first and second modes of assessment. Twelve characteristics of these different modes of assessment are listed in *Table 2*. This list is derived from Thorndike and Hagen (1970), and each characteristic for each mode of assessment is related on a five-point scale from '+ +' superior to '- - ' inferior.

Table 2. Characteristics of different types of questions or tasks

| Characteristics | Task or problem | Essay questions | Short answer questions | Objective questions |
|---|---|---|---|---|
| 1. Measure ability to solve novel problems | + + | + + | + | + + |
| 2. Measure ability to organize knowledge | + + | + + | • | - - |
| 3. Measure ability to evaluate or synthesize knowledge | + + | + + | - - | + + |
| 4. Measure originality or creativity | + + | + + | • | - - |
| 5. Measure writing skills | - - | + + | • | - - |
| 6. Isolate specific cognitive skills from general skills of performance and writing | - - | - - | - | + + |
| 7. Diagnose areas of weakness | - - | -- | + | + + |
| 8. Sample widely content of instruction | - - | - - | + | + + |
| 9. Freed from opportunities to guess answers | + + | + + | + + | - - |
| 10. Scored consistently between makers | - - | - - | - | + + |
| 11. Scored quickly or by machine | - - | - - | - | + + |
| 12. Construct high quality questions in a short time | + | + | + | - - |

*Source:* Thorndike; Hagen. 1970.

++ = superior. + = slight advantage. • = neutral.
- = slight disadvantage. - - = inferior.

In general, it should be noted, classroom teachers have difficulty in constructing high quality objective test items. However, high quality essay questions and performance tasks or mathematical problems are also difficult and require considerable time to construct. *Table 2* indicates clearly that each mode of questioning has its advantages and its disadvantages.

*Criticisms of school assessments*

Criticisms of the use of school assessments are commonly advanced on the grounds of equity. Students from high status backgrounds are seen by teachers to be well-behaved, highly capable of expressing themselves in performance, discussion and writing, and well able to maintain effort because of home support. Moreover, they are highly motivated and assisted on performance tasks by the home to succeed, and well-versed in the use of empowerment skills which they employ to their advantage in the school and classroom. As a consequence, marks assigned to them by teachers are said to be biased in their favour. In a similar way, girls are commonly assigned higher grades than boys (Tyler, 1956). Thus it may be argued that a bias operates against boys in the use of school assessments in coeducational, but not in single-sex schools.

A further argument against the use of continuous school assessments is advanced on the grounds that the student is under sustained pressure for a period of one or two years while assessments are being made. As a consequence, little opportunity is provided for able students to reflect and explore, by undertaking activities, that are associated with intellectual development and scholarly work, but that fall outside the programme of assignments for which grades are given. Such problems are said to be acute in highly competitive situations, where there is considerable pressure to achieve high grades for selection purposes.

*Moderation of school assessments*

The major problem associated with the use of school-assessed marks is that the grades awarded by different schools may not be

comparable. One possibility is to moderate such grades by examination results, which do not count for the individual, but do influence the average level of performance of the school or subject group. The school grades moderated in this way commonly count between 25 and 75 per cent of the total score. This may lead, unless the examination is well designed, to teaching for performance on an examination that has been restricted in scope because of the school assessment programme. Another possibility is to moderate the school assessments using a scholastic aptitude test, as has been used in Australia, so that allowance is made for the general ability of the students in a school group. These different moderation procedures are considered from a technical perspective in *Chapter IV*. Other moderation procedures are used in Sweden and Scotland which are based on a type of quality control with standard tests administered during a school year. In the province of Ontario in Canada, school assessments without moderation have been used in selection for entry into higher education. In Japan, school assessed grades are supplied to universities, but little weight would seem to be assigned to them.

These different moderation procedures probably all work in a satisfactory way in school systems where there is relatively little variation between schools in their levels of achievement. However, in developing countries where there may be marked diversity between schools, both in inputs and achievement outcomes (for example, in Brazil, where the differences between the north-eastern states and the southern states are considerable) such moderation procedures would probably serve to exclude from further education the very able students learning in very low performing schools. Such countries need procedures that could identify future talent, and that would provide talented students with opportunities to undertake further education. It should also be noted that in many developing countries the variation between schools in their level of performance increased greatly from the primary through to the upper secondary grades, so that different approaches might need to be adopted at different levels of schooling.

The quality control procedures used in Sweden for the assigning and moderation of school assessments, probably operate well in a small country, but may prove very difficult to use

effectively in a large country where the logistics of grade recording would seem to create substantial problems.

Likewise, in a small system the training of teachers to make sound assessments of their students' performance that are aligned with examination results can be achieved, to a degree that is not possible in a large school system. Without moderation of school grades, some schools, which have a vested interest in student performance, would not only tend to inflate the marks assigned to their students, and so reduce their discriminating power for selection purposes, but the marks would also tend to be biased through such inflation. Likewise, the use of information on students' personal characteristics, which is supplied in references, and information on participation in extra curricular activities may be used by individual institutions in their selection of students. However, it is difficult to provide this information through a national examination system, and to use it in situations where there is intense competition for a limited number of places as occurs in many developing countries.

## *Form of examination offered*

Differentiation between the two major functions of public examinations, namely, certification and selection has led in some countries to the development of two different forms of examination. In the first form, the examination is defined by the curriculum and the examination seeks to measure the learning of both content and cognitive skills with respect to a specified curriculum. In the second form, the examination is curriculum free, and the examination seeks to measure aptitude to learn under appropriate conditions. More specifically, such a scholastic aptitude test aims to measure verbal and quantitative skills or other special abilities that may not be reflected in a student's assessed level of academic achievement. There is little doubt that the best single index which predicts achievement at a higher stage of education is achievement in the same or a related field at a lower stage. Moreover, this level of prediction far exceeds that obtained with a scholastic aptitude test. For example, McGaw (1976) has reported from a study in Australia that aggregate standardized marks in an external matriculation

examination correlated at a level of 0.6 with success in first year university studies, while performance on a scholastic aptitude test correlated approximately 0.25 with the same criteria.

Experience in the USA indicates that university performance is more highly predicted by a weighted combination of scores on both a scholastic aptitude test and examinations than by either taken alone (Donlon, 1984). The capacity of a scholastic aptitude test to assess ability to learn in the future rather than opportunity provided in the past to learn successfully, suggests that such a test can contribute to selection in situations where the teaching and learning with respect to a specific curriculum may have been deficient. Similarly, in the selection of mature-aged students who apply for entry to higher education from a wide range of educational and employment backgrounds, a scholastic aptitude test can provide useful information on an applicant's capacity to succeed in a particular course.

*Combining school assessments and public examination marks*

A further consideration is the extent to which assessments made within the schools should be used, as an alternative to or combined with performance in an external examination. This issue has been raised in the previous section. There is, however, the question as to whether school assessments should be based on a school examination, conducted at the end of a course, which is similar in nature to a public examination, or whether the school should assess in a continuing programme extended assignments in order to increase the range of performance under survey.

The relative weights given to school assessments and public examination marks in the calculation of a total score, must in part depend on the nature of the school assessments which are used and the area of the curriculum under consideration.

Apart from the increased consistency of a total score, and the greater meaningfulness obtained from measurement under slightly different conditions, there seems relatively little to be gained from the use of school assessments which merely duplicate the measurement of the same skills and content that are obtained from public examinations. A stronger case can be argued for the use of school

assessments that measure a wider range of skills and content than can be elicited in the limited time and format of a public examination, and are more closely aligned to the objectives of the curriculum.

*Alternative forms of public examination*

Public examinations commonly involve only the measurement of those skills that can be assessed in a pencil and paper examination. An oral examination permits, on the one hand, the assessment of a student's capacity to assemble ideas and speak coherently, and on the other hand, the interrogation of a student to clarify a response or to probe depth of understanding, but such examinations are time consuming and costly to administer. In addition, it is uncommon in oral examinations to measure student performance with a high degree of objectivity and consistency although this is not always possible with pencil and paper examination. Likewise, other types of performance, such as investigation in a science laboratory, where practical skills are assessed, or the playing of an instrument in the field of music studies, are costly to measure in a public examination. Consequently, in the design of a public examination it is necessary to consider both the importance of such skills and the contribution of accurate measurement to certification as well as to the prediction of future performance in selection situations.

The danger exists that, where a public examination dominates the work of the schools, and that examination measures only a very limited range of skills, even though the predictive power of the examination might be high, the instruction provided in the schools is unduly and undesirably restricted. This may occur to such an extent that the instruction is detrimental to the long term intellectual and other development of students. Moreover, it may not provide them with the full range of competencies and cognitive skills that they could need at later stages in their education and their lives.

## *Design and mode of public examinations*

Several modes of obtaining student responses are available from which one or more might be chosen to improve both the meaningfulness and consistency of assessment in a public examination. Three aspects of assessment would seem to be important (see *Table 2*). First, there is the need to sample adequately from the full range of content of the curriculum which the examination seeks to asses. The time available for a public examination is necessarily limited, and a large number of items, each of which takes only a short time for a response, permits a more effective sampling of the content of a specific curriculum. Secondly, there is a need to sample adequately across the range of cognitive skills or processes associated with a specific curriculum. Thus to test merely knowledge of content, or computation in a mathematics examination, and not to assess problem solving skills would seem incomplete and inadequate. Thirdly, there is a need to measure student performance with a high degree of objectivity and accuracy. The term 'objectivity' implies that different markers as well as the same marker working on different occasions would assign the same mark to a particular test item or piece of work. This is of particular significance in situations where large numbers of students are tested, and a correspondingly large number of markers are involved in assigning scores to student work.

### *Need for objectivity*

The need for objectivity and adequate sampling has led to the widespread use of multiple-choice and short answer test items in public examinations, particularly in the USA. A multiple-choice, true-false, or matching test item can be assessed with a high degree of objectivity and consistency, and such items are commonly knows as 'objective test items'. A constructed response test item, while eliminating the chance that an item is answered correctly by guessing, is more difficult in general to score with complete objectivity, except in cases where a numerical response is required. Such test items which demand a carefully prepared list of responses that are considered to be correct, are generally referred to as

'quasi-objective test items'. However, in situations where large numbers of students are tested, it is both time-consuming and costly to hand mark constructed response test items. It is even more difficult to mark such items where different score values are assigned to different levels of response.

Objective test items can be answered on mark-sense sheets that can be accurately marked in a short period of time by an optical scanner, and the scores assigned automatically and stored on a computer file. This procedure greatly facilitates the operation of marking student responses in public examinations. While it is possible to train students to respond with clearly written numbers to a specific style, or to write or type in a way that can be read with an optical scanner, the use of optical scanners in the scoring of quasi-objective test items is not common. It would seem likely that in the decades ahead more widespread use will be made of computers in public examining, so that a student responds using a lap-top computer in an examination room, possibly to an examination paper presented on a computer screen. The students' responses could then be more readily scored, and quasi-objective test items could be processed automatically, with a high degree of objectivity. Nevertheless, the cost of such equipment is likely to preclude its use on grounds of equity, since students from lower socio-economic status backgrounds are less likely to have had extensive access to such equipment.

*Need to improve the accuracy of ratings*

Alternative modes of test items widely used in public examinations involve questions that require a response in an extended answer or essay format. Such answers demand assessment involving rating the essay on an extended scale. Much can be done to improve the accuracy of ratings (Guilford, 1954), but maintaining consistency in essay marking remains to some extent problematic. Greater consistency can be achieved by double or treble marking, but this is very costly in large public examinations (Donlon, 1984). As a consequence, double marking, in such examinations, is generally restricted to consistency checks of scoring, to the

investigation of border-line cases, and to the examination of marker reliability.

*Some important choices*

While an essay necessarily tests whether a student can provide a clear and coherent written statement in response to a question (see *Table 2*), the time taken to answer inevitably restricts the range of content or cognitive skills that can be sampled. The advantage that an essay type of response has over a multiple-choice response lies in the writing of a clear and coherent extended answer under examination conditions. Such an essay-type answer generally demands the selection, assembling and presentation of ideas in a connected sequence. Thus, the value of essay-type examinations rests on the importance attached to the students' capacity to exhibit and use effectively such skills. An essay question does not demand other cognitive skills of analysis, synthesis and evaluation, that cannot be tested by a multiple-choice question, except in so far as a student is not required to develop and present an argued case in responding to a multiple-choice question..

*Taxonomies of educational objectives*

It is necessary to recognize that sole reliance on multiple-choice tests and indeed on pencil and paper examinations in the summative assessment of education outcomes can be in the long-term deleterious. While greater costs are involved in the use of other forms of assessments, their substantial educational benefits have to be considered. Furthermore, a major problem is encountered in the use of multiple-choice questions arising from the difficulty that is encountered in the construction of high quality test items which assess cognitive processes other than knowledge and understanding (see *Table 2*). This obstacle has led in some countries to the employment of test items in public examinations that are of poor quality, and which can be strongly criticized with respect to their educational value.

As mentioned above, there is an ever present danger in educational testing to employ items or questions that only require

knowledge in order to provide an answer that is scored correct or given full marks. This danger applies equally to multiple-choice tests as to extended answer tests and even to essay-type examinations. To combat this tendency Tyler, both in the Eight-Year Study and in the programme of the University of Chicago Examinations Center, began to construct multiple-choice test items as well as essay questions that assessed cognitive skills other than knowledge. Gradually, a taxonomy of educational objectives was developed that was subsequently published in a tentative form (Bloom, 1956) but was of such value that it has been successively reprinted until over one million copies have been sold. The Bloom taxonomy has been translated into many languages and has had a highly significant impact on educational practice in the field of testing and examinations (Anderson and Sosniak, 1994). This taxonomy identifies six major categories of: *knowledge*, *comprehension*, *application*, *analysis*, *synthesis* and *evaluation*, with the five categories other than knowledge being considered to be hierarchically ordered as cognitive skills. Kreitzer and Madaus (1993) have reviewed research that tests for the hierarchical structure of the Bloom taxonomy. However, the problem remains that multiple-choice test items that assess synthesis and evaluation are difficult to construct and such skills are rarely tested. Likewise, in essay questions it proves difficult to set questions that demand the use of application, analysis, synthesis and evaluation. They are also difficult to mark when such cognitive skills are required.

De Landsheere (1977) has documented the many different taxonomies that have been developed for use in assessing educational outcomes. However a further taxonomy that is of particular value in the construction of multiple-choice items and mathematical and scientific problems is the *Structure of Learning Objectives* (SOLO) Taxonomy (Biggs and Collis, 1982). This classification of learning objectives builds upon the Piagetian stages of cognitive development to identify five levels of skill, namely: prestructural, unistructural, multistructural, relational and extended abstract stages. This taxonomy would also appear to correspond to a classifactory scheme, employed by the US Educational Testing Service in its programmes of test construction, that uses as a basis of classification one, two and three operations to respond correctly to a test

item or to solve a problem. The SOLO category of 'multistructural' would seem to employ single operations in a chained sequence, while the next highest level 'relational' would seem to involve operating at a second stage on two or more products formed at the first stage, whether they are formed in a chain or not.

Both the Bloom and the SOLO taxonomies imply that instructional practice should be appropriately directed towards the development of higher level cognitive skills and extended cognitive operations. If testing is restricted merely to knowledge, then teachers are tempted to teach for the memorization of knowledge rather than seeking to promote cognitive development, by skilful questioning and structured learning experiences. In the construction of any particular test, or examination it would seem desirable that consideration should be given to the design of the instrument to be used, to the balance of the different modes of questioning and to the different cognitive skills and levels of operation required across the different segments of the instrument and across the instrument as a whole. The principles involved in such work are well presented in *The Handbook of formative and summative evaluation* by Bloom and his colleagues (1971). Examination centres also need to work closely with the curriculum developers to ensure that the grid used in the development of an examination is fully consistent with the design of the curriculum.

## *Procedures of curriculum planning and development*

In *Chapter I Figure 1* presents in diagrammatic form the interrelations between assessment, curriculum and instruction. The importance of these interrelations cannot be ignored in the design of national examinations. As suggested in the previous section, the nature of the assessment instrument employed in a national examination can strongly influence the manner in which instruction occurs as well as the objectives of the curriculum. However, when a syllabus is prepared for a public examination, it is common for only the content to be specified in detail. It is generally considered to be the professional responsibility of the teacher to determine how instruction should take place. Under such circumstances, it is only over time in response to particular types of questions that demand

higher cognitive skills and extended operations, that teachers modify their instructional practices. Moreover, without such pressure from the examinations which are set, instruction sometimes reverts to the mere acquisition of knowledge, generally in rote form and thoroughly memorized.

It would seem highly desirable that where bodies are established for the conduct of public examinations, they should either have the dual functions of curriculum development and assessment or they should be able to interact fully with the persons who develop the objectives and content of the curriculum. This would not only ensure that there is agreement between the two aspects, but also that evaluators and the curriculum developers both acknowledge and jointly promote, through interaction with teachers, beneficial instructional practices. Such steps as involving experienced teachers in both curriculum development, and the planning and marking of public examinations are advantageous, but only if these efforts are directed towards sponsoring beneficial teaching practices. Otherwise, a narrowing of the curriculum, of instruction and of assessment may take place, with close agreement between all three aspects, but with detrimental long-term effects on the intellectual development of students. It is of course possible that because examination centres employ technical and statistical specialists alongside management specialists, they may ignore the detrimental backwash effects of examinations. Moreover, where examinations agencies see their role as separated from the educational system and as auditors, they tend to regard the classroom and the learning that takes place within it as someone else's responsibility.

*Examinations can sponsor change*

Over a long period it has been common among teacher and liberal educators to challenge the use of public examinations on the grounds that they are necessarily detrimental to high quality teaching. However, such an approach ignores the fact that teachers are commonly resistant to change, teaching as they were taught decades earlier. It also overlooks the way in which public examinations can be used to foster and support change.

Heyneman (1987, p. 16) provides a salutary example from Kenya, where:

> "... considerable attention is given by the Ministry of Education to explaining the thought patterns behind wrong answers. Because the incentives for passing examinations in Kenya are so strong, such explanations are quickly integrated into classroom teaching. Test feedback mechanisms may be the most efficacious means that educational managers have to improve the quality of instruction."

It is clearly important to consider how an examinations agency is not only established, but also how it is financed, and how its functions are defined.

## *Administrative problems in examinations agencies*

There are three commonly used models for the establishment of examinations agencies. First, an agency may be set up as an instrument of the government, that has responsibility for education in the region in which the agency operates. Secondly, an agency may be set up by the major users of the results reported from an examination. These are commonly institutions of higher eduction that are involved in the selection of applicants who seek entry to the institutions and take the examination for this specific end. Under these circumstances, the selection function dominates. Thirdly, it is possible for all parties that are 'stakeholders' in the examination process – the teachers, educational administrators, employers, institutions of higher education, the government and the wider community – to be fully and equally represented on the board of the examinations agency. This helps to ensure that the needs and interests of all parties are taken into consideration in the planning and development of the examinations.

Historically, the second model was the dominant one and was generally appropriate where the major purpose of the public examinations was for selection into universities. However, as secondary education expanded, public examinations also served a certification function. Furthermore, the grade levels at which

examinations were conducted moved downwards from the pre-university level, and the public examinations, and thus the universities, established control over secondary schooling. Where there were many stakeholders in the examinations, including both public and private schools, boards with joint representation were generally formed. However, where the major stakeholder was the government school system, it was perhaps inevitable that the examinations agency should have tended to become an instrument of government. The shortcoming associated with an examinations agency being a governmental instrumentality, is that such an agency can be greatly influenced by political pressure. Moreover, an agency run by universities is dominated by university interests which are not necessarily the same as those of the schools. Only an agency controlled by a board comprised of the representatives of the many stakeholders remains genuinely free to conduct its own affairs. As a consequence, there is much to be said in favour of this third model.

*Financing examinations agencies*

An important condition for independence and autonomy in decision-making, which. in addition to appropriate legislation, is likely to lead to a greater maintenance of professional standards in the conduct of a public examination, is an independence of fiscal resources. This independence of fiscal resources is achieved by permitting the examinations agency to charge fees from candidates who take the examinations. The income from these fees remains with the agency and is used to carry out the work of the agency. It is necessary, however, for the level of fees to be large enough to enable the examinations agency to develop its research capacity, and to establish and maintain appropriate technical standards. In addition, the fees charged must not be so large that educational opportunities are restricted for some potential candidates by an inability to pay the necessary fees. It may, thus, be necessary for the examinations agency to receive a subvention from governmental sources which together with a moderate level of fees would permit the agency to conduct its affairs in an autonomous way.

*One or more agencies*

In England and the USA, more than one examinations agency is available for selection by schools to administer public examinations to students. While the existence of multiple agencies has the advantage that competition helps with the maintenance of professional standards, there are other consequent problems of comparability between the results reported by two or more agencies. Although such problems are not insurmountable, comparability of results is not readily achieved (see Pidgeon, 1967). Generally, small countries are unable to support diversity in examinations agencies, and in fairly large countries some diversity is not only possible, but for the efficiency of conduct of the public examinations, it may also be desirable. Multiple agencies would seem to make necessary and desirable curricular changes very much harder to implement. The introduction of curricular change requires considerable expenditure of books and materials, and commonly some retraining of teachers to teach a new curriculum. To fund this expenditure in schools some moneys which would need to be provided from governmental sources, would seem to be necessary. The interrelations between the curriculum developers, the agency that conducts the public examinations, and the teachers who provide instruction in schools indicates that no one aspect can proceed satisfactorily without fruitful interaction with and support from the other two. Flexibility can be provided through the use of item banking procedures which are considered in *Chapter IV*.

## *Equipment required by an examinations agency*

To maintain high standards of technical efficiency, particularly as the size of the operation expands, it is necessary for an examinations agency to have certain major items of equipment. Three major items are probably essential, namely, (1) an optical scanner, (2) a computer, and (3) off-set printing equipment. Much of the important work of an examinations agency is undertaken during limited time periods, with a high degree of accuracy demanded, and under pressure. Consequently, it is necessary for the examinations agency to have back-up equipment or sole access to back-up

equipment in times of emergency. The back-up equipment should be identical to or interchangeable with the main equipment, so that all operations could be readily transferred at a time of emergency. Moreover, the replacement equipment must be capable of being dedicated to the back-up role at specific critical time periods. There is inevitably the danger that expensive equipment, unless shared, may lie idle for long periods of time. Appropriate plans need to be carefully prepared to meet the constraints of very heavy demands on equipment for relatively short periods, and perhaps only one such period each year.

*Optical mark readers*

An optical mark reader or optical scanner permits mark-sensed answer sheets to be employed for the scoring of multiple-choice questions. The main requirements of an optical mark reader are not the volume of answer sheets that can be read per minute, but rather the reliability of the equipment, the accuracy of the settings, and the sensitivity of the reading equipment. The hand processing of mark-sensed answer sheets, or the hand checking of the sheets for erroneous readings is tedious and costly.

*Computers*

An all-purpose computer (or, for example, IBM 486 computers in a network) is essential equipment for an examinations agency. However, the use of the computer depends greatly on the computing skills and technical competence of the staff working with the equipment. It is, nevertheless, not the size of the computer that is important, but the ease with which it can be expanded to cope with an increased volume of work as well as to increase certain aspects of its storage to process large amounts of data in a relatively routine way.

The selection and training of staff to work with both an optical mark reader and a computer is critical for the success of an examinations agency. Where once a large and seemingly powerful mainframe computer was demanded for the work of an examinations agency, it is now possible to achieve much with a personal

computer, or with personal computers linked together, provided the staff have the skills to operate and maintain this computing equipment effectively. Thus, an examinations agency needs not only a carefully prepared plan to acquire and use appropriate computing equipment, but also a plan to maintain and develop the skills of essential staff, who are likely to be highly mobile and perhaps frequently replaced.

*Printing and publishing equipment*

Good printing equipment is also necessary for a well-organized examinations agency. While it may be necessary for the large numbers of examination papers to be printed by a large commercial printer, problems of security may demand that all printing is carried out within the agency. There are nevertheless large numbers of reports, discussion documents and possibly trial versions of examination papers that are best typed and printed by the staff of the agency. Desk-top publishing equipment (for example, a Macintosh computer) is today so versatile, that skilled staff can prepare 'camera ready' copy relatively easily. Moreover, facilities are now available for the drawing of high quality and detailed diagrams using a desk-top computer, provided the staff have the necessary skills. The danger is that skilled staff have to work under immense pressure for short periods and remain not fully utilized for long periods. Careful planning of a programme of operations on an annual basis can help to reduce such problems.

## *Procedures employed in the conduct of national examinations*

The preparation, production, distribution, administration, collection, processing, compilation, verification and reporting of results form a tightly specified sequence and require careful planning, if the conduct of a national examination is to be successfully accomplished. A brief comment is made below on each of the above steps in the sequence.

*Preparation*

The first task is the analysis of the curriculum for the content and skills or processes to be tested. The setting of an examination paper involves the construction of test items and questions to meet carefully specified requirements that are shown to be fully consistent with the defined curriculum. Commonly, a small team works on the setting of an examination paper. This is especially necessary if multiple-choice test items are to be used, because multiple-choice test items of high quality are difficult to construct. It is equally difficult to prepare a sound and detailed scoring plan for essay questions. Consequently, discussion amongst the members of a team, some of whom have had considerable experience teaching students at the level of schooling under survey, helps to identify shortcomings in examination questions, and to check that the questions conform to a syllabus and test the desired skills.

Where possible, it is desirable to submit an examination, particularly one containing multiple-choice or other objective test items, to a trial. Where it is not possible to trial the test items prior to an examination, it is essential that the items should be considered thoroughly by an independent panel of assessors, who were not involved in the development of the examination items. However, if an examination paper carries a heavy burden, such that very large numbers sit for the examination, or if considerable financial rewards are made available to sizeable numbers as a result of performance, field testing of the items is necessary, even if, for purposes of security, the trial must be carried out in another country. The meaningfulness and strength of an examination must generally be investigated in a research programme, which is carried out after the completion of the administration of the examination and the processing of results.

*Production and security*

The printing of the examination paper is an important step, because it must be accomplished without error. The elimination of typographical errors is achieved only if the proof reading of the galley proofs is undertaken by a skilled proof-reader who also

understands the subject content being tested by the examination. Printing and proof reading and subsequent storing of examination papers must commonly be carried out with total security. The more at stake in an examination the more rigorous must be the security precautions. Possibly the weakest link in the chain of events for which total security must be maintained is immediately prior to the examination, when the printed papers are held in the offices of a large number of examination centres. It is also particularly important that the sitting by students for an examination should occur at exactly the same time in all examination centres throughout a region or even a country. Differences in times of administration make possible the transfer of examination papers between students, even across large distances.

Some examination papers are so expensive to produce, such as those assessing scholastic aptitude, that it is necessary for all copies of an examination paper to be held in tight security even after the examination. The ease with which a photocopy of an examination paper can be made, allows breaches of security to occur without examination papers being reported as missing from a carefully numbered sequence. While examination papers are rarely used on a second occasion without change, if trends in student performance over time are to be examined, then some items must be replicated on two or more occasions. Any consideration of such items prior to taking an examination must serve to distort the comparisons over time, and this no doubt is a major reason why so few studies of performance trends over time have been undertaken. It should also be noted that it is sometimes convenient to field test items by including them as trial items in an examination or testing programme. While students are unable to identify such items they are neither used in obtaining students' scores nor in making comparisons over time.

Another problem which arises from lack of security after an examination has been administered is that a lost or copied examination paper facilitates coaching for the examination. While practice under examination conditions and systematic preparation is both necessary and desirable, the effects of extensive coaching are questionable. It is, however, not uncommon with a purpose of fairness to all, to make readily available those past examination

papers that are not costly to produce. There would seem little danger in providing practice and limited coaching using past examination papers, if these papers are themselves educationally sound and reflect the defined objectives of a course. Alternatively, it may be of value to release a sample of items from a bank of examination questions for widespread use, while withholding other items for later use in the investigation of change over time and the maintenance of standards.

*Administration*

It is essential if comparisons are to be made between the performance of students taking an examination at more than one examination centre, that the conditions for the administration of the examination must be completely standardized. This requires that all students, wherever located, take the examination under identical conditions. Not only must the time allowed for students to sit the examination, including time for preliminary reading of the paper, be carefully controlled, but the conditions in the venue in which the examination is conducted must also be controlled. Moreover, the instructions given to the students must be controlled, either by all instructions being printed on the examination paper, or by the invigilators being required to read a carefully prepared and printed statement that is thus the same for all students.

*Processing*

The goal in marking completed examination papers is to reduce the variation both between different markers and between the same marker on different occasions. This objectivity in the marking of examination papers can be fully achieved through machine scoring procedures. For example, student responses to multiple-choice questions can be keyed into computer files, independently verified and the responses scored by computer. Alternatively, optical mark reader answer sheets may be used, and the scoring carried out by an optical mark reader. Nevertheless, problems can arise in the use of optical mark readers from extraneous marks being placed on the answer sheets, or from careless marking by students of the answer

sheets. Optical reading procedures are possible with extended response questions, and some work has been done to develop the computer scoring of essays. While such computer scoring operations are moderately successful, there is insufficient confidence in their use for them to be employed in scoring national examination papers.

Although complete consistency cannot be achieved through the rating or marking of student responses, greater objectivity must always be sought through the procedures employed. In general, carefully formulated marking schemes must be used in the rating of students' answers to mathematical problems and extended response questions. Systematic checks also need to be applied to ensure that raters maintain consistency over time, as well as consistency between raters in conforming to the specified marking scheme. While it is always possible to compensate for systematic bias between raters, or to revert to the check marking of borderline cases, or cases where there are discontinuities in a student's pattern of performance, it would seem preferable to attempt to specify in detail the scoring system to be employed, and to train markers to adhere to it.

Some controversy would appear to exist in the marking of essays that involve creative writing, because the style of responses may differ so greatly that an effective scoring system is difficult to construct. Under such circumstances, global judgement scoring is commonly advocated on a reduced scale, perhaps with only five scale points, with relative rapid reading of an essay or sample of writing, and then to require that all responses are double or treble marked by independent scorers. The two or three scores for each response are subsequently added together. Even in such situations it is necessary for standards of performance to be specified in detail and for markers to be required to undergo thorough training, with systematic checking of their ratings to ensure that comparability is maintained. Donlon (1984) briefly describes the different approaches to this problem that have been investigated by the Common Entrance Examination Board in the USA since the mid-1940s in the quest for an adequate level of consistency in the scoring of performance in achievement tests in English.

*Compilation of results*

In the compilation of results, there is little reason in the 1990s not to make the fullest possible use of computers. The scores assigned to a student need to be entered into a computer and fully verified, if not already scored from an optical scoring operation. Systematic checks for wild-codes, discrepant values and missing data should also be made. Whereas, in survey research, attempts are made to estimate missing scores, this would seem inappropriate in national examinations. In very exceptional circumstances it may be of value to recognise that such estimation can be done meaningfully. For example, if the whole or a part of a student's answers to an examination is known to have been lost at an examination centre, then it is commonly assumed that the student must be credited with maximum performance. It may, however, be more satisfactory to employ score estimation procedures.

*Verification of results*

Random errors are extremely rare in computer processing. Systematic errors, which arise as a result of programming errors, or perhaps brief power failure when a computer is running for a long period, must always be anticipated. A series of checks must be applied at successive stages to guard against the occurrence of such errors. It would also seem desirable not only to apply internal checks at every stage, but also to apply checks back against the raw data recorded on student answer sheets using recognized quality control procedures. However, the aim in the computer processing of examination results, unlike quality control in manufacturing, is completely error free products. Major errors in computing can have disastrous consequences for an examinations centre, particularly since they are commonly only detected after results have been released. Even minor errors, while perhaps readily provided for, serve to destroy confidence in all the operations employed.

*Maintenance of high standards of operation*

High standards of conduct of national examinations can only be achieved as a result of the professional training of staff and efficient management of the examinations agency. While financial support for the examinations agency is not the only factor associated with efficient and effective operation, it is at least necessary to have adequate funding. It is also important that an examinations agency is not susceptible to political influence. Ultimately, the quality of operation of an examinations agency resides in the capacity of the agency to pursue professional objectives in an autonomous way. It is not essential to have competition between two or more examinations agencies in a particular country. However, in England and the USA where this is known to have occurred, it would seem to have had beneficial results. Whereas in other countries, where only one agency exists, there is a tendency for the establishment of highly bureaucratic structures, that can become closed, inefficient and very conservative in outlook.

## *Reporting of results from examinations*

There are many stakeholders in national examination systems, and the reporting of results must meet the needs of all concerned. Nevertheless, the initial reporting of results is to the candidates who sat for the examination. However, the schools and institutions that prepared the candidates and the teachers within those schools are not only interested in the results of their students, but they themselves are judged by the performances of their students. Where the results of an examination are used for selection, then it is also necessary that information on the results is passed on to the institutions making the selection. Likewise, where the results of an examination are used for certification, then employees and parents also need to be provided with information that would assist with the interpretation of results in individual cases.

Furthermore, the wider community, including politicians have concern for the:

- effectiveness of a national examination in the maintenance of educational standards;
- in matters of equity for groups in the society:
- examination's unintended influence, on the lives of students, whether beneficial or detrimental;
- effective teaching and learning in schools;
- costs of conducting the examination; and (6) and the examination's contribution to the cost-effectiveness of the educational system.

Each stakeholder has different interests, and the needs of each must be met in such a way that confidence in the system is maintained.

*Reporting to students*

The results of student performance must be released in the form of grades that can be readily comprehended. The critical decision as to what level of performance constitutes a pass is, however, difficult to determine. Standards of performance are not readily defined in any absolute way. The most than can be provided from the distribution of scores on an examination is a relative level of performance. If, however, the frame of reference for relative performance is gradually shifting as a result of an increased proportion of the age group taking an examination as commonly occurs, then average standards of achievement must be expected to decline. Thus, while relative performance retains certain meanings, a shift is difficult to detect with respect to absolute standards. Experienced examiners consider that they can detect a shift in absolute standards of performance. This may or may not be true. However, if a decline occurs as a result of a higher proportion of the age cohort taking an examination, then it is necessary to ask whether standards should be maintained and a higher proportion of candidates failed. Alternatively, should a fixed proportion of the candidature, say 80 per cent or 75 per cent, or 67 per cent always pass? Moreover, it must be recognized that different subjects attract candidatures of different levels of ability or scholastic aptitude. For example, physics can be shown to attract more able students than biology (Keeves, 1992). Consequently, it is necessary to consider whether the proportion of students passing a physics examination

should compensate for this difference, irrespective of the quality of teaching in the schools.

Such issues are not readily resolved. Indeed, appropriate data have not been examined on some of these issues, and probably are not available for consideration outside the USA. Even in the USA there are major problems of interpretation, because such examinations are conducted on a competitive basis by two or more agencies. The meaning of grades, both on a particular occasion and over time, is clearly unknown, except with respect to the relative standing of students. Likewise, scores assigned on a scale from 0 to 100, as is commonly employed have the appearance of greater accuracy and perhaps meaning than grades on a five point scale. Except in so far as a hundred-point scale provides greater discrimination than a five-point scale in a relative sense, both scales have no absolute meaning either between subjects or over time.

The issue of equating or scaling of scores between subject fields is considered in *Chapter IV*. While such scaling provides for difference in the quality of candidatures between different subjects and in the quality of the assessment instrument being used, the need for such procedures is neither widely recognized, nor is the manner in which the adjustments to scores might be made, widely understood. As a consequence, the recording of grades alongside adjusted scores, or even unadjusted scores together with adjusted scores generally proves to be confusing when results are reported to students. Claims of lack of fairness are common where such reporting is attempted. Likewise moderation is sometimes employed to equate scores obtained across different schools or institutions and is also considered in *Chapter IV*.

Whatever the procedures employed for calculating the relative performance of candidates across schools and across subject areas, it is essential to recognize that a programme of public relations needs to be developed. Such a programme would not only present results but would also build up an understanding of the form of presentation, and the reasons for the use of that form in the presentation of results. A variety of different approaches must be used to convey to different audiences the information that they might wish to know. To conceal or withhold information, on the grounds that the issues cannot be understood by a non-technical

audience can be very damaging in the long-term, although perhaps leading to little debate in the short-term. Thus, it is necessary for examinations agencies to include in their annual budgets sufficient funding to permit the development of a strong public relations programme for the release of information on the scoring and reporting procedures employed.

### *Reporting to schools*

There is a further area of reporting that is of particular importance, namely the reporting back to schools of information that would lead to the improvement of the quality of instruction, as well as information that would lead to the effective revision of the curriculum. Brief comments by examiners on students' answers to particular questions, with judgements commonly made in terms of absolute standards of performance, would seem inadequate. More information is clearly required if the standards of performance and of teaching and learning are to be raised. As a consequence, consideration should be given to the public release by the examiners of worked solutions to problems, together with the marking scheme employed, as well as information on the rating of answers to extended response questions, again with sample answers and with justification of the ratings assigned. With multiple-choice test items diagnostic information on the errors made in the choice of distractors as well as the knowledge, understanding and skills required to select the correct alternative can be of considerable value to teachers and their students. Such information can greatly assist in advancing the teaching and learning that takes place in schools.

## *Problems in the conduct of examinations*

### *Undesirable practices*

In countries where there is strong competition for entry to university, parental, and peer group pressure to succeed lead to certain problems and undesirable practices in the conduct of public examinations. Such practices include the stealing of copies of

examination papers from a printing house, or from the office of an examination centre prior to the administration of an examination. In addition, they may involve substitution of a candidate in the examination hall by a person who is able to perform extremely well on the examination. Alternatively, cheating during an examination by using concealed notes may also occur. Direct copying by one person of another's work is difficult to accomplish in an examination hall where desks are well spaced and where there is proper supervision of students. However, where school assessments form a major component of the scoring procedure employed, assignments may be written by a person other than the student submitting the assignment. Copying large sections of an assignment from an undisclosed source may also occur. Some of these undesirable practices are extremely difficult to detect, and recent investigations of such practices suggest that they may be more widespread than is commonly assumed. It is argued that external examinations, administered under standardized conditions, by paid invigilators who are not the teachers of the students, largely eliminate these undesirable practices and thus lead to providing results of greater value. However, strong rules concerned with breach of accepted conduct are needed, and should be enforced by a well-managed agency that also applies monitoring procedures to detect irregularities.

*Coaching and private tutoring*

In Japan, private tutoring is very widespread and clearly reflects the strongly competitive nature of Japanese schooling and the procedures of selection into institutions of higher education. It is difficult to separate those practices of tutoring that are educationally beneficial and which assist a student in the mastery of subject content for examination purposes and those that are detrimental. One of the reservations advanced against the use of a scholastic aptitude test, as part of an examination programme is that performance on such tests is susceptible to coaching. However, it is clear that limited practice on these tests is advantageous and improves performance, but only to a limited extent. Consequently,

all candidates should be given an opportunity to practice the taking of such a test.

Pike (1978) has argued that coaching and special preparation for an examination can affect a student's scores in three different ways. *First*, it can develop the skills and abilities that the examination seeks to measure. *Secondly*, it can improve a student's familiarity with the content, format and specific types of tasks that the examination tests. *Thirdly*, it can raise a student's level of confidence and efficiency in taking the examination. By raising performance in all three respects, particularly the first, which can be said to correspond to high quality instruction, such coaching is beneficial. If coaching is directed towards the memorization of prepared answers to standard questions that are similar or identical to the ones asked in an examination, then the educational value of such coaching must be questioned. Likewise, if the coaching is primarily a commercial enterprise, that trades on the anxiety of a student to achieve well in a particular examination, then the practice should be strongly discouraged. Some of these undesirable practices can be greatly reduced or eliminated by the release of past examination papers, preferably with notes and comment as to what are considered by the examiners to be answers of high quality, and what are the common errors in answering the examination questions. This enables the industrious student to learn from the past and to advance such a student's knowledge of the subject. Likewise, the provision of opportunities for the students to obtain practice in the taking of an examination without payment of additional money, as is commonly involved in coaching or tutoring, enables improvement in performance to occur in ways that are fair to all students.

## *Equity issues in the conduct of examinations*

The question of the fairness of an examination is an important consideration. Furthermore, one of the advantages of public examinations over reliance solely on school assessments, is that, by the impersonal nature of a public examination that is conducted by a central agency, well removed from individual schools, the external examination system is both fair and is seen to be fair. Likewise, the use of achievement examination results without

consideration of other factors for selecting students to enter universities and colleges is commonly considered to be fair and just. Nevertheless, certain equity issues remain which are associated with public examinations.

*Gender issues*

In *Table 1* information on retention rates for 1984 at the terminal year of secondary schooling is recorded. In all countries for which data were available, except Papua New Guinea, the retention rates for girls exceeded those for boys, although there was no difference between the sexes in England and Sweden in retention rates at the pre-university level of schooling. Admittedly, the data available on developing countries in this table are scarce, but it could be argued that the changes that have taken place during recent times in the composition of upper secondary school classrooms which involve greater participation in education, have not only corrected some of the gender bias that formerly existed, but have also gone beyond a position of balance until girls occupy the dominant place, particularly in such countries as Finland and Hungary.

Further differences are observed with respect to participation and performance in particular subjects. In general, at the terminal secondary schools stage, in 18 out of 19 of the countries listed in *Table 1*, boys outnumber girls in physics classrooms. In chemistry, in 13 out of 19 countries there are more boys than girls in the chemistry classrooms and laboratories at the pre-university stage of schooling. However, in biology the reverse is found and it is relatively uncommon for more boys than girls to be studying the subject during the final year of schooling. However, in all three science subjects, the performance of boys, almost without exception, exceeds that of girls at this level, in spite of the selection bias introduced by differences between the sexes in participation in the study of the science subjects. It is clearly desirable that the changes in both participation and performance in these subjects should be monitored over time (see Keeves, 1992; Postlethwaite and Wiley, 1992).

It is also generally found that boys perform better than girls on multiple-choice tests, problem-solving tasks, and in the field of

mathematics at the middle and upper secondary school levels. Girls, however, perform better than boys on essay tests, in written composition and foreign language learning, and are generally assigned higher grades in school-based assessments. As a consequence the extent to which a public examination includes multiple-choice test items as opposed to essay questions or gives greater weight to internal school assessments as opposed to external examinations conducted over a limited time period, would appear to influence the overall performance differences between girls and boys. It is not clear whether or not such differences really constitute bias in the design of a public examination. Nevertheless, it is necessary for consideration to be given to this issue and for an ongoing programme of research to monitor such effects, as they are likely to change over time.

Attitudinal factors, which differ between the sexes, have also been shown to be related to performance in public examinations. Such factors include motivation in academic work and liking of school, where girls hold more favourable attitudes, and confidence in success, where boys have more positive attitudes. Whether efforts should be supported to raise boys' levels of motivation, liking of school and participation in education at the terminal secondary school level, and girls confidence in success remains an unresolved issue. However, during the 1980s and 1990s in many countries programmes were introduced and have continued to operate in order to compensate for gender effects in schooling that work to the disadvantage of girls.

*Social class issues*

Evidence from the studies of educational achievement carried out by the *International Association for the Evaluation of Educational Achievement* (IEA) (Husén, 1967; Comber and Keeves, 1973; Keeves 1992) show that in the subject areas of mathematics and science at all levels of schooling and in all countries, there are significant differences between social groups, identified in terms of the status of the occupations of the fathers of the students, in the average levels of achievement of the groups.

At the terminal secondary school level, there are differences in retention rates between social class groups, with student from higher status homes staying longer at school, and with only students of greater ability from lower status homes remaining. The observed differences in levels of achievement between the social class groups, persist although they are reduced in size as a result of differences in retention rates. Similar results are to be expected in achievement in public examinations.

*Firstly*, it is sometimes argued that such differences are a consequence of the fact that public examinations place a premium on writing skills, which involve verbal ability, and that such skills and abilities are more highly developed among students from homes in the higher social status groups. However, while verbal ability is a factor of some importance in later life, it is only one of the many factors involved and bias is introduced if undue importance is given to this factor.

*Secondly*, it is argued that higher status homes provide an educational environment that fosters a higher level of motivation to succeed in public examinations, and to gain selection for higher education. This enhanced motivation accounts for a higher level of performance in public examinations by students from higher status homes that may not be sustained in subsequent education.

*Thirdly*, it is sometimes contended that higher levels of achievement among students from higher status homes are a consequence of their attendance at fee-paying schools which are said to provide a better education than do the public schools (for example, Coleman; Hoffer, 1987). Comparisons of performance between different types of schools in many countries would seem to lend support to these claims. However, the students in such fee-paying schools are commonly drawn from more educative and higher status homes. Attempts have been made to allow for the status of the home and other factors in controlling between school differences in achievement for the effects of the home. However, the methods that have been employed have generally been highly questionable, because of the influence of 'aggregation bias' when data at the student level are aggregated to the school level. The statistical problem is a complex one and some developments of

significance which allow for multilevel effects have recently occurred (Raudenbush and Willms, 1991; McPherson, 1993).

Greater social equity can probably be achieved by making available a greater number and a wider range of opportunities for higher education, and by policies of more open access. Nevertheless, such opportunities are generally only taken up by more highly motivated students.

*Ethnic and racial issues*

Mathews (1985) in the United Kingdom, and Breland (1979) and Durán (1983) in the USA have examined the issues of whether public examinations are biased against certain racial and ethnic groups. Such issues are of importance in countries where there are sizeable ethnic and racial minority groups. Nevertheless, it is not always that students from an ethnic minority group achieve at a lower level, or are under-represented in higher education as a result of a lower level of performance in a public examination.

There seems little doubt that where students are taught and are examined in a language that is different from their mother tongue they will be at a disadvantage compared with those students whose mother tongue is the language of instruction. As a consequence, it must be recognized that the need to unify a country through the use of a common language, which is necessarily used as the language of instruction in schools, gives rise to some bias against students from racial and ethnic groups within the country. Moreover, students are also at a disadvantage who are migrants to a country, or who speak a language different from the language of instruction, in their homes. The degree to which talent is lost, and the extent to which compensatory actions might be taken are unclear. However, such questions might well be considered as to whether or not the language used in public examinations minimizes difficulties of comprehension for minority students, and whether or not examiners are sensitive to the difficulties some students may have in the use of language in responding to examinations.

## *Conclusion*

From the discussion presented in this chapter several ideas emerge as being of importance in the design and conduct of public examinations:

- There is need for highly trained personnel to administer and to work in a national examinations agency.
- There is need to maintain a programme of research and development regarding all aspects of the conduct of public examinations.
- It is necessary for a public examinations agency to be adequately funded to ensure that programmes of training and research are sustained.
- There is need to have access to both optical mark reading and computing equipment. However, the costs of the training of staff to use such equipment effectively are of greater consequence than the costs of purchase of such equipment.
- There is need for the staff of an examinations agency to maintain close liaison with those responsible for curriculum planning and development, and with those who teach in schools.
- Consideration needs to be given in the design and development of national examinations to the balance between different item types, whether multiple-choice, essay questions, or performance tasks. In addition, consideration must be given to processes and cognitive skills being assessed in an examination, and whether the range can be increased with profit, through the use of school-based assessments.

*Finally*, issues of equity and social justice are of increasing importance in many countries, both highly developed and developing countries alike, and continuing investigation into these issues needs to be maintained, so that informed decisions on such issues can be made.

# IV. Some technical issues

In this chapter, five technical issues are considered in some detail. While in the space available it is not possible to examine these issues in depth, they are of such importance for the design, procedures and reporting of national examinations that they should not be overlooked or ignored. The discussions of these issues are grouped together into one chapter because they are inter-related, rather than to suggest or imply that some readers might well pass over this chapter in a first reading of this booklet.

Unless these issues are addressed in the planning of national examinations, particularly as expansion in participation to the terminal stage of secondary schooling takes place, there is the danger that a well established examination system will fail to adapt to changing context and conditions.

As a consequence, an examination system may be prematurely abandoned on ideological grounds rather than those that contribute to the raising of educational standards. This can be seen to have occurred in a small number of the countries as considered in *Chapter II.*

The five issues treated in some detail in this chapter are:

(i) modelling levels of achievement with changing retention rates;

(ii) scaling to allow for differences in candidatures between subjects;

(iii) moderation of school-assessed marks between schools;
(iv) use of item response theory (IRT) scales and comparisons of achievement over time; and
(v) use of item response theory in item banking.

## *Modelling levels of school achievement with changing retention rates*

Circumstances arise where retention rates and participation rates differ both between countries and within a particular country over time, and it is necessary to make comparisons. In such comparisons, allowance must be made for the effects of differences of selection.

McIntosh (1959) first proposed that the truncated normal distribution might be employed to model the effects of selection on performance on the 11+ Examination in the United Kingdom. This model was developed further by Walker (1967) to account for the differences between the mean mathematics scores at the terminal secondary school level in the First IEA Mathematics Study conducted by the *International Association for the Evaluation of Educational Achievement* (IEA).

Although the IEA assessment studies do not involve directly examination performance, the model they employ can be extended to the consideration of differences in performance in public examinations over time. The first basic assumption underlying the use of the model is that within each country in successive years there is the same distribution of ability in the total age group, which influences performance in an examination subject.

Secondly, there is the assumption that the differences in the means and variances found in performance over time in a public examination in that subject are the result of the selection procedures that led to changes in retention or participation rates over time.

The simplest specific assumptions that relate to the model are:

- the scores each year on a subject examination would be normally distributed over the total age group, if all in the age group had taken the examination;

- these hypothetical distributions of scores are identical for each year; and

- the students examined each year are the best performing students in the age group of that year and in that subject area.

On this basis it is possible to calculate the expected mean scores and variances on the subject examination of the groups of students who form the selected part of the age group tested each year, if the proportion of the tested group to the total age group is known.

The formulae for these calculations are derived from the truncated normal distribution shown in *Figure 2.*

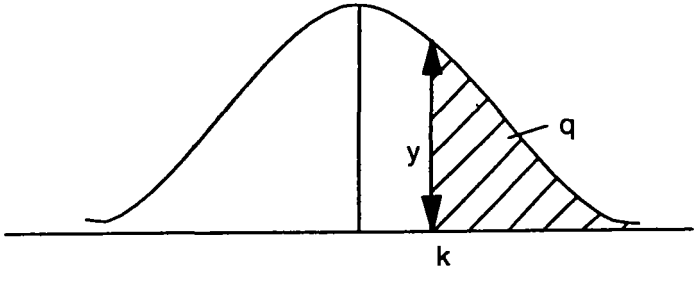



Figure 2. The truncated normal distribution

$$\text{mean} = \frac{y}{q}$$

$$\text{variance} = 1 - \frac{y}{q}\left(\frac{y}{q-k}\right)$$

where $q$ = proportion of age group selected for testing;

$y$ = ordinate of the normal curve at the cut-off point;

and $k$ = point of cut-off.

One shortcoming of the model is that it assumes that retention at school and participation in the study of the subject under survey has operated to select the best performing students in that subject. This assumption might be violated because subject choice could be influenced by requirements laid down for entry into institutions of higher education, or by patterns of study demanded to ensure that students maintain balance across different fields of study. The model can be modified to assume a correlation $r$ between the variable employed to select students, referred to as the retentivity variable, and the subject examination scores. While such correlations are not known, they are likely to remain relatively constant over time within a country, or for discontinuities to be clearly recognizable as being the result of changes in policy or practice. In the testing of the model, plausible values may be selected and assessed against data for goodness of fit. If $r$, the correlation is assumed to be constant over time, the means and variances may be derived as:

$$\text{mean} = \frac{ry}{q}$$

$$\text{variance} = \frac{r^2 y}{q}\left(\frac{y}{q} - k\right)$$

This model was tested by Walker (1967) and was found to be consistent with the achievement data on the IEA Mathematics tests measured in 1964.

Adams (1984) also tested the model with data collected in the Australian Capital Territory on a scholastic aptitude test over a period of seven years, for male and female students separately, with the assumptions that the population correlations between the retentivity variable and the scholastic aptitude test scores were, for males, and for females. In both cases, convincing levels of fit were observed. The truncated normal distribution model of achievement is also consistent with the inverse linear relationships recorded in the First and Second IEA Science Studies (see Comber; Keeves, 1973; Keeves, 1992; Postlethwaite; Wiley, 1992) between achievement in science at the terminal secondary school level and retention rates and participation rates in the study of science.

While the effects of improved teaching might be expected to lower the correlation between the retentivity variable and subject examination scores, there are unlikely to be marked changes in such effects over time. Substantial changes are likely to occur if an additional year of schooling is provided, or if major programmes are introduced to support financially able students who would otherwise have dropped out from school. Moreover, stronger relationships are likely to be recorded if a total subject score, for all examination subjects taken by a student as used for selection into higher education, is employed than if a single subject examination score is used in the testing of the model.

In discussions of standards of achievement in national examinations over time, this model of school achievement, which considers the effects of selection, would appear to be of considerable value. It provides a simple and coherent explanation for what many staff of higher educational institutions have observed over time, namely, a decline in the entry performance of students as the educational system expands through an increase in retention rates. The only ways to compensate for such a decline in standards of achievement would seem to be to provide a longer period of study or more intensive study in higher educational institutions or by substantial advances in instruction at the school level.

## *Allowance for differences in candidatures between subjects*

It is common practice in the selection of candidates for entry into higher education to combine together the scores obtained in different subjects to form an aggregate or total score. Moreover, it is common to permit students to study different combinations of subjects, so that the total scores are based on subject scores which are awarded to groups of students taking the different subjects (referred to here as different student candidatures), where the groups differ markedly in size and also in the average and spread of abilities of the students. The problems raised by differences in candidatures were of little significance when retention rates were low, for example, less than 5 per cent of the age group, and the range of subjects studied was small. However, in those countries where retention rates have greatly increased, it has also become necessary to increase the range of subjects studied and examined, and as a consequence the problems of lack of comparability between candidatures have increased.

Relatively little systematic work has been done to illustrate the magnitude of this problem, because of the difficulties associated with the administration of ability tests to large groups of students at the terminal secondary school level. Keeves (1992) shows using a word knowledge test as a surrogate for a verbal aptitude test, and a computational skills test as a substitute for a quantitative ability test, the extent of the problem with large national samples of students from several countries. Students taking physics are clearly a more able group than those taking biology at the terminal secondary school level. Such differences arise from the manner in which different school subjects are taught and in the demands made in teaching on competence in quantitative reasoning in the study of the physical sciences, or on verbal ability in the learning of classical and foreign languages, or in some countries, on logical argument in the reading of philosophy. Unless some allowance is made by scaling for these differences in ability of the candidatures taking different subjects there is a serious lack of fairness in combining together scores for different subjects to obtain a total score. The different statistical scaling procedures described below

are employed to make allowance for differences in subject candidatures.

*Use of a standard test*

In its simplest form this procedure involves the adjustment of scores on the *i*th achievement test Yi using a standard test X. Two assumptions are made, namely: (1) the joint distribution of the standard test and the achievement test scores is bivariate normal, and the marginal distributions are normal; and (2) the standard test has a significant correlation *r* with the achievement test. The following linear equation can be employed in the process of adjustment.

$$T_{ij} = X_{i.} + r \frac{S_{xi}}{S_{yi}} ( Y_{ij} - Y_{i.} ) \qquad \textit{Equation 4.1}$$

where
- $T_{ij}$ is the adjusted score of student j taking achievement test i;
- $Y_{ij}$ is the score of student j on achievement test i;
- $Y_{i.}$ is the mean score on achievement test i;
- $X_{i.}$ is the mean score on the standard test of the group taking achievement test i;
- $S_{yi}$ is the standard deviation on the achievement test i; and
- $S_{xi}$ is the standard deviation on the standard test X of the group taking achievement test i.

An extension of this procedure is employed in the USA for the scaling of College Board subject examination scores using the two standard tests of verbal scholastic aptitude ($SAT_v$) and mathematics scholastic aptitude ($SAT_m$). In formulating the linear equation for the adjustment of scores, the partial regression coefficients are employed for predicting achievement using $SAT_v$, and holding $SAT_m$ constant, and using $SAT_m$, and holding $SAT_v$ constant. The major problem with this procedure is that performances in different subject areas do not correlate equally with the two standard tests

that are employed ($SAT_v$ and $SAT_m$). The greater the correlation between the achievement test variable and the standard variables, the greater the values of the adjustments made (Donlon, 1984).

In three states of Australia a standard test, which is a test of scholastic aptitude, has been used, each with slight variations, and with the assumption that *r* the correlation between the achievement test i and the standard test is unity (r = 1) This avoids the injustice to some subjects that results from the relatively low correlations of the achievement test scores in those subjects with the standard test scores. Equation 4.1 is restated as:

$$T_{ij} = X_{i.} + \frac{S_{xi}}{S_{yi}} (Y_{ij} - Y_{i.}) \qquad \textit{Equation 4.2}$$

where the meaning of the coefficients is the same as in Equation 4.1. This procedure does not lead to the candidates taking achievement test *i* being no longer ranked in the same order, but that the means and standard deviations of the adjusted scores differ from those of the achievement test scores.

Equipercentile scaling could also be used if the evidence suggested that the assumption of a bivariate normal distribution did not hold or either the achievement test scores or the standard test scores departed markedly from a normal distribution. These approaches to adjustment of scores for differences in candidatures have sometimes been rejected because it is contended that the scores obtained in different subject examinations do not correlate equally with scores on the standard variables that might be employed, and as a consequence the standard scale so formed is inappropriate.

*Use of an aggregate score*

The most obvious characteristic of differences between candidatures taking a particular subject in a public examination is the achievement of the students in the other subjects that they took on the same occasion. This procedure involves the calculation of an aggregate score on a specified number of other subjects which are

taken by a student. For each subject group these aggregate scores can be averaged for the group, and the mean can be rescaled using these values of the mean and standard deviations of the averaged aggregate scores for the group. This is a simple linear rescaling operation which can be expressed symbolically as follows:

$$T_{ij} = \sum_{k=1}^{m} \sum_{j=1}^{n_i} \frac{X_{jk}}{mn_ik} + \frac{(Y_{ij} - Y_{i.})}{S_{yi}} S_x \qquad \textit{Equation 4.3}$$

The symbols used are similar to those employed in Equation 4.1, except that $Y_{ij}$ is the score for student $j$ on achievement test $i$, and $X_{jk}$ is the score for student $j$ on other achievement test $k$, and where there are ni students in each subject group, and a maximum of $m$ subjects is used in the aggregate. $S_x$ is the standard deviation of the averaged aggregated achievement test scores.

The adjusted score $T_{ij}$ replaces $X_{jk}$ for each subject in turn and the operations can be repeated iteratively until the values of $T_{ij}$ converge. In situations where a large number of students are taking each subject and with considerable overlap in candidatures between subjects, the iterative procedure yields stable results. However, where there are high correlations between the scores for particular subjects and low correlations between scores for other subjects, those groups of students taking subjects that are highly correlated, such as the mathematics and science subjects, have a greater spread of scores ($S_x$), than are estimated for those groups of students taking subjects that do not correlate highly with each other. As a consequence, this procedure may be seen to be biased in favour of certain student groups. Thus, boys studying mathematics and the sciences receive adjustments to their scores that place some at a very high level of achievement, while their counterparts at the other tail of the mathematics and science score distributions have adjusted scores that place them at a very low level of achievement.

An alternative approach is to make no allowance for spread of scores. Only the common scores of the subject groups of students on their other subjects is taken into account in the calculation of

adjustments which are readily computed and yield a unique solution for the adjusted score matrix, and this has no call on iterative procedures. While this approach avoids adjustments that would appear to have a gender bias because boys take subjects that are highly correlated more often than girls do, it is clearly not as accurate as other procedures because it does not take into consideration the shape or spread of the score distributions.

The problems of apparent gender bias also arise in the use of a standard test for adjusting scores, in so far as the mathematics and science scores are commonly more highly correlated with the standard test scores, than are scores in the humanities subjects. This occurs in part, because scores in mathematics and science are obtained with greater consistency and less error of measurement than are scores in such subjects as art and drama.

## *Moderation of school-assessed marks between schools*

In situations where schools assess student performance in a subject and the school assessments are used either alone or in combination with examination scores in a subject to yield a combined subject score, it is necessary to make some allowance for the fact that schools and teachers have different standards in their marking of assessments both between schools and between teachers within schools. The procedures which are employed to adjust scores for such lack of comparability between institutions are referred to as moderation. The moderation procedures employed in such situations may involve either the use of statistically based adjustments or the use of judgements based on the sampling of student work.

It should be noted that moderation may be required on four counts. *First*, there are differences between schools in the characteristics of students who attend them. *Secondly*, there are differences between markers, both between and within schools in the mean level of scores, the spread of scores and the shapes of the score distributions that they assign to students. *Thirdly*, there may be differences between the courses of instruction that the students have studied, and as a consequence of the assessment procedures employed, differences between schools. *Fourthly*, there may be

differences between schools in the quality of the students that are attracted to different subjects within each school. As a consequence, it is not simply a problem of making adjustments between schools that are averaged across all subject areas. It is also necessary to make adjustments across schools at the subject level.

Moderation procedures have been developed to equate levels of achievement for the marks assigned within a school. As previously discussed, the main advantages of using school assessments are that they provide greater:

- scope for flexibility of syllabus and teaching methods;
- opportunities for matching examinations with individual school and local needs;
- latitude to measure outcomes other than those that can be assessed in a pencil and paper examination, such as a portfolio of student work (Mitchell, 1992); and
- freedom to measure student performance throughout the extended period of teaching and learning in preparation for a summative assessment, rather than in only a short time at the conclusion of a course.

Several approaches have been developed, depending on whether or not the school assessed marks are used in isolation, or whether they are used in combination with examination marks.

*Assessments used in combination with examination marks*

In these situations, whatever the weights given to the school assessed marks relative to the examination marks, the examination serves as a standard test, which is common across all schools and that permits an adjustment to be made using differences in levels of performance between schools. The linear equation given in Equation 4.1 is used to make this adjustment, which takes into account the differences in spread of scores, but not differences in the shapes of the score distributions. If these are of consequence, and the school student groups are large enough, equipercentile scaling procedures may be employed. The procedure which employs the linear equation would seem robust enough to be used with school subject groups as small as ten students. For smaller student subject groups it may be advantageous for two or more

schools to combine their assessment procedures and also to join together in the moderation process. This combining of subject groups within two or more schools, can be very effective.

*Assessments not used in combination with examination marks*

Where school assessments are not used in combination with examination marks a standard test, such as a scholastic aptitude test, may be employed not only to adjust scores for differences across subjects but also to adjust scores for differences within the same subject across courses. Normally the adjustments are carried out in two stages. First, subject group assessments are adjusted employing Equation 4.1 and performance on the standard test, a scholastic aptitude test that can be used across subjects. Adjusted scores in a specified number of subjects are then calculated for each student and these total scores are added and averaged at the school level, while student performances on the standard test are also added and averaged at the school level for the same group of students. Further adjustments are made again using Equation 4.1 across schools. Where school subject groups are small in size this two-stage procedure would seem preferable to a one-stage procedure which may be readily applied with stable results where the subject groups within schools are relatively large.

The criticisms advanced against the use of this procedure is that the school is ultimately measured and judged in terms of the level and spread of scores on the scholastic aptitude test taken by its students. No allowance is made for whether or not the school might have taught its students well both generally and in a particular subject. In practice, this procedure establishes a scholastic aptitude test scale on which the qualities of the candidatures of different school and subject or course groups are measured. The performances of the groups as groups are adjusted using measured scholastic aptitude. It is thus heavily dependent on the meaningfulness and consistency of the scholastic aptitude test that is used as a basis for the scaling of school and subject groups. Any bias that exists in the aptitude test scores, whether favouring any social class, racial, ethnic or sex groups, is automatically reflected in the

adjusted scores that are located along the standard scale that is based on the scholastic aptitude test.

The use of a scholastic aptitude test avoids the administration of a series of subject-based examinations, which are employed solely for the purposes of moderation and are sometimes referred to as 'generalized achievement tests'. Thus students take only one moderating test and not several such tests. This single test is also used for scaling between subjects in order to bring the scores in different subject areas to a common scale as discussed above. The benefits of not administering a common subject-based examination is that teachers are free to develop their own curricula in a subject field, with its own specific content and methods of instruction. The benefits of bringing the scores in different subjects to a common scale is that students are not constrained in the subjects they choose to study. The different scaling procedures described above give rise to very similar scaled scores for moderately large, but not for small subject groups. The standard test procedure would appear to work well for small subject groups, while the use of an aggregate score does not.

*Moderation using quality control procedures*

The Swedish school system has developed quality control procedures where the school-assessed marks in key subject areas are checked against standard achievement test scores. These tests are administered at specified times during the final three years of schooling. Some of the standard tests employed are merely diagnostic in nature, while those in six subject areas serve directly to maintain the highest possible degree of uniformity in the marking system, without resorting to the statistical adjustment of scores. All marks are categorized as grades on a five-point scale. The distribution of grades for a class subject group are compared with the distribution of grades on the corresponding standard test in each of the six subject areas. If the two mean grades are identical, or if the difference does not exceed $\pm$ 0.2 it is accepted that the subject grades are appropriate. A conference is held within a school of all teachers involved in assessment at the terminal level of schooling, and those assessment grades that do not conform to

expectation in terms of performance or grades on the standard tests are jointly considered at the conference. The purpose of the conference is to make and justify final decisions on the distribution of grades for all subjects taught at the school. The alleged shortcomings of these procedures are:

- the dependence on the normal distribution for assigning grades;
- the failure to give adequate recognition to the high quality of candidatures of subjects such as mathematics and the physical sciences;
- the equal weighting given to all subjects in the curriculum, although only six are monitored; and
- failure to consider the differences between subjects in time given to studying the subject (Lindberg, 1991; Marklund, 1988).

*Judgemental moderation*

In some situations, such as are involved in the moderation of scores or ratings across schools in the fields of art, music and drama, statistical moderation is considered inappropriate because of the lack of a consistent and meaningful standard scale of performance. However, moderation can be achieved by the judgements of standards of performance made by several moderators who assess and rate samples of student work. The undertaking of such judgemental moderation procedures while necessary is both costly and time-consuming.

## *Item response theory scales*

In the previous sections, consideration was given to the location of persons on an achievement test scale, initially for a single subject, and subsequently for a total score based on a multiplicity of subjects. If, however, it were also possible to locate along the same scale, test items ranging from multiple-choice test items to the ratings assigned to student performance on essays, in an appropriate position, and with sufficient accuracy, then these statistical procedures would have transformed the task of measurement of student learning. In the period from the 1960s to the 1990s,

item response theory (IRT) has been developed using several different measurement models, that may be employed under different measurement assumptions which are related to the nature of the items. As a consequence it is now possible, in the measurement of educational achievement, to construct a scale along which both students and items can be located (see Bejar, 1983; Hambleton, Swaminathan, 1985; Hambleton et al., 1991). Such scales have not as yet been widely used in public examinations, so that the examples of their use that are available are of necessity drawn from assessment programmes. Nevertheless, the development of such scales illustrates well, applications that could be made in public examinations in the future.

*National assessment of educational progress scales*

The *National Assessment of Educational Progress* (NAEP) in the USA has employed a modified three-parameter model to construct a scale of reading proficiency. Sampling information, estimated item parameters and information about individual students have been used in the development of a reading scale extending across the ages of schooling from age nine years to 17 years, and by extrapolation beyond these ages and the corresponding Grades from Grade 4 to Grade 12, in order to measure the reading proficiency of USA youth. This scale of reading proficiency can be considered to be based on a hypothetical test with known properties. Once the scale had been developed, several fixed scale levels were anchored and illustrated by specific test items in order to specify in terms of general statements of reading skills what students at these fixed levels were able or not able to do (Beaton, 1987, pp. 381-390).

The reading proficiency scale, so developed, is an interval scale, and in a sense the scale is also a ratio scale in so far as an estimated score of zero would mean that a student answered no items correctly. However, the zero defined in this way is determined arbitrarily by the specified difficulty parameters of the items, and a zero score does not necessarily imply that a student has no proficiency in reading.

The five fixed scale levels were selected on this scale ranging from zero to 500 with score values: rudimentary (150), basic (200), intermediate (250), adept (300), and advanced (350), to identify what a student could or could not perform with respect to this scale of reading proficiency. By defining the five levels of performance along an interval scale and identifying items to describe each level, it was possible to provide an account of what both an individual student and a group of students could achieve on a reading proficiency test, and thus measure growth. As a consequence, the NAEP reading scale was not only an interval scale but also criterion-referenced with respect to the five levels of performance. The construction of this scale of reading proficiency was a major step forward in the development of a meaningful scale for the measurement of achievement in a particular subject area (see Lapointe, 1985). These approaches to scaling have been applied and extended in the development of achievement test scales for reporting results in the area of mathematics (Dossey et al. 1988) and science (Mullis and Jenkins, 1988).

*The IEA science achievement scale*

A similar item response theory model, the one-parameter or Rasch model, was used in the Second IEA Science Study to compare the achievement in science of students at three age levels (the 10-year-old, the 14-year-old, and the terminal secondary school levels) on two occasions (1970-71 and 1983-84) in 10 countries. The fixed point on this scale, which is shown in *Figure 3*, was set at 500 for the mean difficulty of the items in the test administered to 14-year-old students in 1983-84. The natural unit on this scale was set at 100 scale points. Five levels of achievement were defined on this scale: basic (250), elementary (375), intermediate (500), advanced (625) and specialist (750). The scale value for no measurable knowledge was below the zero point, and the scale value for slight knowledge was above the zero point (Keeves; Schleicher, 1992). The constraints imposed on the test items using the Rasch model are more rigorous, although the conditions under which items satisfy the constraints of the model remain ambiguous.

However, the results of applying these constraints, should lead to more accurate, meaningful and more robust measurement.

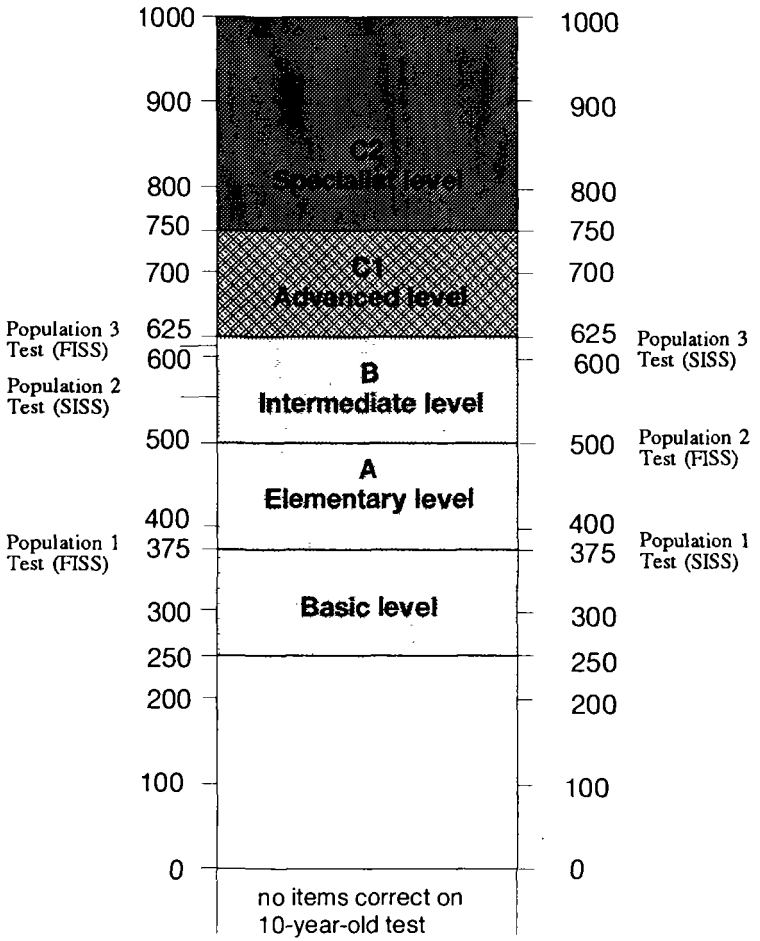



Figure 3. The science achievement scale

Nevertheless, further work remains to be done to test fully the strength of these procedures and to resolve certain measurement and statistical problems that appear to arise when measurement extends across time, across different curricular contexts, and across a wide span of age and grade levels. Nevertheless, evidence from research indicates that the Rasch scale is more robust than the scale based on the three parameter model employed by the National Assessment of Education Progress.

If these problems can be resolved and the conditions laid down under which it is meaningful to construct such scales, whether the one-, two-, or three-parameter IRT model, or another kindred model is used, then the issues addressed in Sections 2 and 3 of this chapter would be transformed, as would the tasks of certification and selection which are required of public examinations.

## *Use of item response theory in banking*

It is not uncommon for a school system to wish to develop a public examination in which there is perhaps 80 per cent of the content tested, and thus of the test items, that is common to all schools or all regions, and 20 per cent of the content and items that is unique to each school or region. Item-response theory provides a scaling procedure by means of which the optional or alternative component can be scaled against the common or core component in a way that is independent of both the sample of questions and the sample of students taking the examination. Problems may arise if the questions used in the optional component have not been field tested to ensure that they have the characteristics which permit scaling to be carried out in a satisfactory way. This has led to the proposal that a large bank of appropriate items should be developed to be employed in the construction of both the core and the optional components of an examination. From this bank of items, questions for a particular examination could be selected, and if the measurement properties of the items were known, estimates could be made of the characteristics of the examination, which would be subsequently confirmed or modified after administration. Indonesia is proceeding with the development of a large item bank that would be used in the conduct of public examinations in which each region

would include an alternative 20 per cent of questions to conform to the optional component in its curriculum (Umar, 1994). Such a strategy in the conduct of national examinations might be appropriate in situations where regional examinations centres have been established such as the West African Examinations Council and the Caribbean Examinations Council. In this way, high quality examinations could be developed that were not only comparable across countries in a region, but also provided for optional national components.

This chapter has considered certain technical issues that have arisen in the conduct of public examinations. The full implications of these issues have not been addressed, and the concluding chapter of this monograph seeks to draw out some of the consequences that flow from advances in measurement and statistical analysis.

# V. National examinations – what is and what might be

The changing nature of the school population, greatly influenced by rising retention rates and the changing composition by social class, sex, age, and racial or ethnic mix of the students making up that population in each country, imposes growing demands on those responsible for the conduct of national examinations. At the same time, advances in educational measurement, and in the technology available to conduct large public examinations efficiently and effectively, have ensured that these expanding demands can be met. Within each country, the responses made evolve over time, and are influenced by contextual, economic, ideological and political factors. These are not immutable, and in each country a long range plan must be developed, but with sufficient flexibility to enable modifications to occur without disruption to the examination system, as conditions, statistical techniques and technological equipment change.

Thus, it might be envisaged that within the foreseeable future some applicants for university entry might take a selection examination at a computer terminal, under conditions of security imposed not by the examination hall, but by monitoring with a video-recording camera. Alternatively, an applicant for employment might sit for a certification examination, again working at a computer terminal in the office of the employer, that is of such a nature that knowledge and skills gained on the job rather than in the classroom are assessed. Nevertheless, immense problems continue to exist in the rapidly expanding school systems of many developing countries, and consideration must be given to these problems, but with a

recognition that unknown and unpredictable changes will inevitably occur in the future.

## *Problems of selection*

The problems involved in selection for entry into many institutions of higher education dominate the concerns about public examinations that exist in many parts of the world. Richter (1987) argues that the social approach in Sweden, the bureaucratic approach of West Germany and the political approach in France, which were tried in the 1960s and 1970s, all failed. There was, however, during this period, structural reform in higher education, an opening up of the traditional highly selective secondary schools, and a maintenance, in all three countries, of open entry into universities for high school graduates. These developments had the goals of raising the levels of qualification and the size of the professionally trained workforce, and of providing greater equality of educational opportunity. Nevertheless, over time it became clear that expansion did not automatically lead to greater equality of opportunity, that open access resulted in an oversupply of personnel in the most rewarding professions, and that more highly qualified manpower did not necessarily create more jobs.

Similar problems occur in many countries and solutions are found that constantly require change as the demand for higher education continues to grow. This imposes a continuing burden on the public purse, and private institutions arise even in those countries where traditionally there has been a state run system of higher education. If the numbers of places available are restricted, selection procedures are essential. They may raise the quality of intake, but they may reduce equality of opportunity, and they do not necessarily meet the requirements of the professional labour market. It is, however, important that the selection procedures are seen to be fair, that the instruments of selection employed are sufficiently discriminating, and that the effects of the public examinations involved in the selection process are not detrimental to the work of the schools.

## *Problems of certification*

National examinations, in addition to providing a selection service, also provide certification for those who seek employment and entry into the labour force. These two functions of national examinations are not necessarily compatible. Where a small proportion of the age group is involved the problems arising from the dual functions of public examinations are readily contained. If, however, a very high proportion of the age group takes a public examination, then the examination has primarily a certification purpose. This situation would seem to demand the development of a two stage procedure, where certification is provided at a first stage, and selection for entry into specific institutions of higher education follows from separate examinations held at a second stage. This occurs in the United States in some states.

This lack of compatibility between the dual functions of public examinations is largely a consequence of the need for either a norm-referenced or alternatively a criterion-referenced approach in measurement. A selection examination demands a high level of discrimination near the cut-off point, while a certification examination seeks to measure the level of competence of a candidate in specified areas. In the past, these two functions, while being seen to be different, were met largely by the use of a norm-referenced examination. The norm-referenced type of examination was expected to provide information on standards of competence, which was more appropriately the function of a criterion-referenced examination.

In certain parts of the world, a movement was set in train in the late 1980s by which certification for employment was to be provided in terms of competence with respect to specified criteria. This movement gained support in Australia, United Kingdom, New Zealand and the USA. Without advances in the measurement of the specific competencies, and without the extension of assessment procedures to measure competencies at different levels, which are necessary as prior conditions for entry into a further stage of education or training, the two functions of national examinations would seem to remain incompatible. If, however, the use of item response theory scaling procedures could be developed to measure

levels of performance on an extended scale, then the operations of both selection and certification could be transformed. A suitably constructed scale of performance could enable those students who had attained a desired level of competence to be certified as having performed at a specified standard. If, however, selection were also required, those candidates performing higher on the scale could be selected. The work carried out by the National Assessment of Educational Progress and IEA indicates that both functions might well be accomplished in the future by using scales of achievement.

## *Problems of distortion*

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. The national examination raises a target for which the earlier years provide the foundation learning, and the later years the more highly focused learning of the knowledge and skills measured by the examination. As a consequence, it is important that the examination should have beneficial effects on the teaching and learning that takes place at all earlier stages of schooling. This is only achieved if those responsible for the design of a national examination maintain close links with those responsible for the planning of the curriculum and the specification of the curriculum objectives, as well as with those who provide instruction in the schools.

Examiners cannot operate independently of the curriculum developers and the teachers. Moreover, it would seem unsatisfactory that examiners should merely be subject-matter experts. They also need to be mindful of well balanced pedagogical principles and the best in teaching and learning practice in order that they should maximize the beneficial influence of the examinations on classroom teaching. It is important that examinations should test more than knowledge, because the mere testing of knowledge both through certain types of objective or multiple-choice test items, or through the memorization of prepared answers or dictated notes as responses to essay questions, must be considered to have deleterious effects on the nature and quality of both teaching and learning.

Examiners and examinations agencies must be aware of such back-wash effects. They must not see their role as auditors or evaluators, but as one that is shared with curriculum developers and teachers, and one in which some responsibility is held for the quality of classroom teaching.

## *Training of staff for an examinations agency*

In the training of staff for an examinations agency, there is a wide range of skills that must be possessed by different members of the staff. Not only is specialist knowledge of educational measurement required, alongside a high level of knowledge and understanding in specific subject areas, but staff also need to be informed on the psychological and pedagogical basis of learning and teaching, as well as recent developments in these areas. There is a need for psychometricians and statisticians, for survey research specialists, educational research workers, publishing experts, art and graphics designers, computer programmers, people who can communicate effectively with the wider public, and production and distribution managers, as well as cost accountants. Above all, it is necessary for an examinations agency to contain members of staff in key positions who have strong negotiating skills, because the successful conduct of national examinations demands negotiation with a wide range of clients and stakeholders.

Each year and in some situations, several times a year, the conduct of an examination must take place without error, and with due regard for the substantial effects of the operations on the lives of individuals. The many different operations involved in the conduct of national examinations demand highly trained staff. Moreover, these staff are working in areas where developments are taking place, sometimes quite rapidly. This requires on-going training programmes for all members of the staff of an examinations agency.

## *Need for a sustained programme of research*

The successful conduct of an examinations system over an extended period of time also requires a commitment by the

examinations agency and its staff to a programme of research. Many problems need to be addressed in addition to those concerned with the strength and meaningfulness of the examination instruments, and the consistency of the marking procedures employed. Such problems include the effects of the examination on the schools, the issues of equity discussed in *Chapter III*, and the accuracy in prediction of alternative selection procedures. Further research issues are concerned with the consequences of changes in retention rates at school and in participation rates in particular subject areas. In addition, it is of value for research to be conducted into the level of achievement in the country with respect to levels of achievement in other countries. There are demands for up-to-date comparative information on the achievement outcomes of education across countries as well as for information on differences in the inputs to education between countries. Without such information, developed and developing countries alike have no standards by which to examine their educational efforts, nor to judge the effectiveness of their educational programmes.

## *Conclusion*

The conduct of national examinations has passed through two major phases and is entering a third phase of reorientation. In the first phase, the purposes of the examination were primarily those of *selection*, since places in secondary and higher education were extremely limited. With the expansion of secondary education, national examinations also undertook the functions of *certification* together with those of selection, and as participation in secondary education increased, tension developed between the two functions. As movement occurs in many countries – at least the most advanced – towards universal secondary education, with 12 years of schooling for a very high proportion of the age groups, and as programmes of recurrent and lifelong education become widespread, new approaches are needed that can be used well into the twenty-first century. There are signs that with the advances of item-response theory in educational measurement, it may be possible to develop more appropriate procedures for reporting the results of different forms of national examinations. Eckstein; Noah (1992,

1993) have considered some of these dilemmas encountered in the conduct of public examinations. However, the resolution of these problems must be viewed from a perspective which is influenced by the phase of educational orientation that exists within a particular country. While it is necessary to ensure that examinations agencies operate efficiently, with due regard to the costs involved, and making optimal use of new technology, it is also essential that a national examination system should meet the emerging demands of the society it serves.

# References


Adams, R.J. 1984. *Sex bias in ASAT?* Hawthorn, Victoria, ACER.

Anderson, L.W.; Sosniak, L. 1994. *Bloom taxonomy. NSSE Year Book.* Chicago, National Society for the Study of Education.

Bernstein, B. 1990. "The social construction of pedagogic discourse". In *the structuring of pedagogic discourse.* Vol. 4. Class, Codes, and Control. London, Routledge.

Bottani, N.; Duchene, C.; Tuijnman, A. 1992. *Education at a glance.* The OECD indicators, OECD: Paris.

Beaton, A.E. (ed.) 1987. *The NAEP 1983-84 technical report – implementing the new design.* Princeton, New Jersey, Educational Testing Service, National Assessment of Educational Progress (NAEP).

Bejar, I.I. 1983. *Achievement testing: recent advances.* Beverly Hills, California, Sage.

Biggs, J.B.; Collis, K.F. 1982. *Evaluating the quality of learning.* The SOLO taxonomy. New York, Academic Press.

Bloom, B.S. (Ed.) 1956. *Taxonomy of educational objectives: Handbook 1.* London, Longman.

Bloom, B.S.; Hastings, T.J.; Madaus, G.F. 1971. *The Handbook on Formative and Summative Evaluation of Student Learning.* New York, McGraw-Hill.

Breland, H.M. 1979. *Population validity and college entrance measures.* College Board Research Monograph, No. 8., New York, College Entrance Examination Board.

Coleman, J.S; Hoffer T. 1987. *Public and private high schools: the impact of communities.* New York, Basic.

Comber, L.C.; Keeves, J.P. 1973. *Science education in nineteen countries.* Stockholm, Almqvist and Wiksell, New York: John Wiley.

De Landsheere, V. 1977. "On defining educational objectives". *Evaluation in Education: International Progress* Vol. 1, No. 2, pp. 104-111

Donlon, T.F. (ed.). 1984. *The College Board Technical Handbook for the scholastic aptitude and achievement tests.* New York, College Entrance Examination Board.

Dossey, J.A.; Mullis, I.V.S.; Lindquist, M.M.; Chambers, D.L. 1988. *The mthematics report card.* Princeton, New Jersey, Educational Testing Service, National Assessment of Educational Progress (NAEP).

Durán, R.P. 1983. *Hispanics' education and background: pedictors of college achievement.* New York, College Entrance Examination Board.

Eckstein, M.A.; Noah, H.J. 1992. *Examinations: comparative and international studies*, Oxford, (United Kingdom) Pergamon.

Eckstein, M.A.; Noah, H.J. 1993. *Secondary school examinations, international perspectives on policy and practice.* New Haven, Yale.

Faure, E. et al. 1970. *Learning to be. The world of education today and tomorrow.* Paris; UNESCO.

Guilford, J.P. 1954. *Psychometric methods* (2nd edn.). New York, McGraw-Hill.

Halls, W.D. 1985. International Baccalaureate. Husén T.; Postlethwaite T.N. (eds.) In *The International Encyclopedia of Education* Vol 5, (1st. edn.) Oxford, Pergamon.

Hambleton, R.K.; Swaminathan, H. 1985. *Item-response theory. Principles and applications.* Boston, Massachusetts, Kluwer-Nijhoff.

Hambleton, R.K.; Swaminathan, H.; Rogers, H.J. 1991. *Fundamentals of item-response theory.* Newbury Park, California, Sage.

Heymeman, S.P. 1987. *Uses of examinations in developing countries: selection, research, and education sector and management.* EDI Seminar Paper Series No. 36. Washington, D.C. The World Bank.

Heyneman, S.P.; Fagerlind, I. (Eds.) 1988. *University examinations and standardized testing. Principles, experience and policy options.* Washington D.C., The World Bank.

Hidano, T. 1988. "Admission to higher education in Japan". In Heyneman, S.P.; Fagerlind, I. (Eds.) 1988. *University examinations and standardized testing. Principles, experience and policy options.* Washington D.C., The World Bank.

Husén, T. (ed.). 1967. *The international study of achievement in mathematics; a comparison of twelve countries.* (2 vols.). Stockholm, Almqvist and Wiksell.

Husén, T. 1979. *The School in question: a comparative study of the school and its future in Western society.* London, Oxford University Press.

Husen, T.; Keeves, J.P. (eds). 1991. *Issues in science education: science competence in a social and ecological context.* Oxford, (United Kingdom), Pergamon.

Keeves, J.P. (ed.). 1992. *The IEA Study of Science III: Changes in science education and achievement: 1970 to 1984.* Oxford, (United Kingdom), Pergamon.

Keeves, J.P.; Schleicher, A. 1992. *Changes in science achievement: 1970-84.* Keeves, J.P. (ed.) op. cit., pp. 263-290.

Kreitzer, A.E.; Madaus, G.F. 1994. "Empirical investigations of the hierarchical structure of the taxonomy". In Anderson, L.W.; Sosniak, L. *Bloom taxonomy. NSSE Year Book.* Chicago, National Society for the Study of Education.

Lapointe, A. 1985. *The rading report card.* Princeton, New Jersey, Educational Testing Service, National Assessment of Educational Progress (NAEP).

Lindberg, Y. 1991. Comments. Husén, T. Keeves, J.P. (eds). In *Issues in science education.* Oxford, (United Kingdom), Pergamon Press. pp. 212-3.

Linn, R.L.; Baker, E.L.; Dunbar, S.B. 1991. Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher.* 20(8).

Lu Zhen. 1988. "A brief introduction to the system of higher school enrollment examinations in China". In Heyneman, S.P. Fägerlind, I. (eds.) *University examinations and standardized testing. Principles, experience and policy options.* Washington D.C., The World Bank.

Madaus, G.F.; Kellaghan, T. 1992. Curriculum evaluation and assessment. [P.W. Jackson. *Handbook of Research on Curriculum*. New York, Macmillan].

Marklund, S. 1988. "Education in Sweden: assessment of student achievement and selection for higher education". In Heyneman, S.P. ; Fägerlind. I. (eds.) *University examinations and standardized testing. Principles, experience and policy options*. Washington D.C., The World Bank.

Mathews, J.C. 1985. *Examinations: a commentary*. London, (United Kingdom) Allen and Unwin.

McGaw, B. 1976. The use of rescaled teacher assessments in selection of students to tertiary study. *Australian Journal of Education*. 21(3), 209-25.

McIntosh, D.M. 1959. *Educational guidance and the pool of ability*. London, University of London Press.

McPherson, A. 1993. Measuring added value in schools. SET: *Research Information for Teachers*. 1(3), 1-4.

Mitchell, R. 1992. *Testing for learning: how new approaches to evaluation can improve American schools*. New York, The Free Press.

Mullis, I.V.S.; Jenkins, L.B. 1988. *The science report card: elements of risk and recovery*. Princeton, New Jersey, Educational Testing Service, National Assessment of Educational Progress (NAEP).

Pidgeon, D.A. 1967. *Achievement in mathematics: a national study of secondary schools*. Slough, Bucks. NFER.

Pike, L.W. 1978. *Short-term instruction, testwiseness and the scholastic aptitude test. A literature review with research recommendations*. College Board Research and Development Report 77-78 No. 2, Princeton, N.J., Educational Testing Service.

Postlethwaite, T.N. 1967. *School organizations and student achievement: a study based on achievement in mathematics in twelve countries*. Stockholm, Almqvist and Wiksell.

Postlethwaite, T.N.; Wiley, D.E. 1992. *The IEA study of science: science achievement in twenty-three countries*. Oxford, Pergamon.

Raudenbush, S.W.; Williams, J.D. 1991. *Pupils, classrooms, and schools: international sudies of schooling from a multilevel perspective*. New York, Academic Press.

Richter, I. 1987. "Selection and reform in higher education in Western Europe". In Za'rour, G.I. *The role of examinations and testing in educational management*. Washington D.C., The World Bank.

Rosier, M.J.; Keeves, J.P. (eds.) 1991. *The IEA study of science I: science education and curricula in twenty-three countries*. Oxford, Pergamon.

Tyler, L.E. 1956, *The psychology of human differences* (2nd edn.). New York, Appleton-Century-Crofts.

Tyler, R.W. 1949. *Basic principles of curriculum and instruction*. (Published in many places and translated into 25 different languages). Chicago, University of Chicago Press.

Thorndike, R.L., Hagen, E.P. 1977. *Measurement and evaluation in psychology and education* (4th edn.). New York, Wiley.

Umar, Y. 1994. Item Banking. [T. Husén, T.N. Postlethwaite (eds.) In *The International Encyclopedia of Education* (2nd edn.). Oxford, Pergamon].

Walker, D.A. 1967. "An attempt to construct a model of the effects of selection". Husén, T. (ed.)1967. *The international study of achievement in mathematics; a comparison of twelve countries.* (2 vols.). Stockholm, Almqvist and Wiksell.

ZA'ROUR, G.I. 1987. *The role of examinations and testing in educational management.* Washington D.C., The World Bank.

# IIEP publications and documents

More than 750 titles on all aspects of educational planning have been, published by the International Institute for Educational Planning. A comprehensive catalogue, giving details of their availability, includes research reports, case studies, seminar documents, training materials, occasional papers and reference books in the following subject categories:

*Economics of education, costs and financing.*

*Manpower and employment.*

*Demographic studies.*

*The location of schools (school map) and sub-national planning.*

*Administration and management.*

*Curriculum development and evaluation.*

*Educational technology.*

*Primary, secondary and higher education.*

*Vocational and technical education.*

*Non-formal, out-of-school, adult and rural education.*

Copies of the catalogue may be obtained from the IIEP Publications Unit on request.

# IIEP publications and documents

More than 750 titles on all aspects of educational planning have been, published by the International Institute for Educational Planning. A comprehensive catalogue, giving details of their availability, includes research reports, case studies, seminar documents, training materials, occasional papers and reference books in the following subject categories:

*Economics of education, costs and financing.*

*Manpower and employment.*

*Demographic studies.*

*The location of schools (school map) and sub-national planning.*

*Administration and management.*

*Curriculum development and evaluation.*

*Educational technology.*

*Primary, secondary and higher education.*

*Vocational and technical education.*

*Non-formal, out-of-school, adult and rural education.*

Copies of the catalogue may be obtained from the IIEP Publications Unit on request.