Int. J. Educational Development, Vol. 7, No. 4, pp. 251–263, 1987
Printed in Great Britain

0738–0593/87 $3.00 + .00
Pergamon Journals Ltd.

USES OF EXAMINATIONS IN DEVELOPING COUNTRIES: SELECTION, RESEARCH, AND EDUCATION SECTOR MANAGEMENT*

STEPHEN P. HEYNEMAN

Education and Training Design Division, Economic Development Institute, The World Bank

Abstract—In September 1984, The Chinese government asked if the Economic Development Institute of the World Bank would be interested in assisting the officials in the Ministry of Education to think through some of the policy options in the field of examinations and standardized testing. In response, in April 1985 a meeting was held in Beijing. Attending the meeting were all officials in charge of examinations at the provincial and national levels, technicians and psychometricians in charge of designing examination items, and senior university officials and planners in the Ministry of Education. Attending from outside the country were the chief executive officers of examination agencies in three OECD countries: from the United States, Robert Solomon (ETS); from Japan, Tadashi Hidano (National Center for University Examinations); and from the United Kingdom, John Reddaway (Cambridge University Examination Syndicate); directors of the National Assessment of Educational Progress and the International Association for the Evaluation of Educational Achievement; and experts on the examination systems in Sweden, Australia and Kenya.

This article summarizes the comments given to the Chinese Government following that meeting. The amount of attention devoted to problems of logistics and economies of scale is perhaps more pertinent to large, heterogeneous nations like China, but many of the comments could be applied to developing countries generally. These comments cover three areas (1) specific testing issues such as aptitude vs achievement tests, multiple choice vs other formats, etc.; (2) management issues within the system of selection such as whether government agencies or universities should make the selection decision, and whether a testing agency should be autonomous from government control; and lastly, (3) the uses of testing to perform necessary research and education sector management functions.

GENERAL BACKGROUND

Why should those concerned with the economics of developing countries pay attention to the problems of educational selection? In a competitive international environment, not choosing one's technical elite from among the brightest citizens can have a grave effect on economic performance. By one estimate, developing countries could improve their Gross National Product per capita by 5% if they were to base leadership upon merit; by another estimate, the economic pay-off to developing countries would be three times more than the pay-off were OECD countries to reduce restrictions on third world exports (Pinera and Selowsky, 1981; Kirmani *et al.*, 1984). Though the magnitude of effect may be debatable, the theory is reasonable.

The theory suggests that certain, but not all, elements of social selection are amenable to policy manipulation. Within the Education Sector there are basically three mechanisms —(1) whether a wide group of citizens enters school; (2) whether they stay in school; and (3) how the few of them chosen to attend university education are selected. This discussion concentrates upon the use of tests to manipulate the third mechanism.

The use of selection tests concerns all developing countries, even those with near universal rates of attendance and grade-to-grade progression. When selection tests are used, educational systems are strongly affected. Selection tests tend to produce relatively tangible results by which to judge quality. Pressures felt on educational systems which use

*Opinions and views in this paper are those of the author and do not necessarily represent the views of the World Bank or any of its affiliated institutions.

selection tests are sometimes broadly 'popular' since, despite technicalities, results can be widely interpreted in the press and by various voluntary associations and interest groups. Selection testing draws attention at one specific time to a single, widely-understood indicator. It holds the school system accountable for results; and it fosters an open and continuing forum on the school system's ability to deliver results parallel to the public's expectations.

To be sure selection tests create individual anxieties, and anxiety does in fact, affect results. Moreover, testing often highlights differences in the quality of educational inputs and learning opportunities among sub-groups of a population. Because of these problems, some progressive developing countries such as Tanzania and Indonesia began to dismantle their national systems of selection tests in the late 1960s and instead began to rely upon university-designed tests or upon selection criteria other than test results. The latter have been used to rectify past injustices to specific sub-groups; to ensure fair geographical representation; to recognize and reward abilities other than the academic. But both replacements have created unforeseen problems.

University-designed tests have turned out to be logistically cumbersome, as the number of universities expanded, and no more reliable as tests. Selection criteria other than testing have not turned out to be free of anxiety nor necessarily more fair. When school systems have relied upon particularistic criteria—political loyalty, family alumnae status, personal wealth, ethnicity, geography of birth—the effect has often been even more pernicious than the abandoned system of centralized testing. Those not selected by these criteria may feel that the choice was made unfairly. This has occurred for instance when the selection choice has been made subjectively on the basis of political attitudes. Resentment and political backlash has also been known to occur when selection choices have been made on the basis of ethnicity. There the damage can be two-sided. Those selected may never feel as though they personally 'deserved' the opportunity; those not selected may feel that the choice was made on an unacceptable basis since the rationale was to account for problems of group representation rather than individual worth.

But not all tests are equally fair; nor are all tests equally efficient. Many countries are shifting back toward using selection tests and this, in turn, has raised new problems. The best way to handle these problems has been to 'manage' them better. This implies that there should be a conscious effort to anticipate dilemmas to which there is no single answer; a careful program to select the specific innovations most likely to fulfill domestic requirements; and a hard nosed attitude of cost containment even though it will require an abrogation of long-held traditions separating testing and research agencies.

The following section lists four testing issues where these dilemmas should be anticipated. On none could simple resolution be expected. The last section lists six areas of testing which are amenable to general recommendations, that is, where policy guidelines can be explicit. This section is followed by a brief summary.

TESTING DILEMMAS

Which model?

Testing is like a manufacturing process in that techniques are rarely invented independently. Though the content of specific examination questions may reflect local culture, there are only a few models from which to draw. Which model is preferable? Which characteristics are exportable?

In many OECD countries educational testing is a sizeable private industry. In the United States and Australia this industry is internally very competitive and is represented by numerous firms, most of them meeting other needs than post-secondary selection. Clients consist of public and private schools, school systems, colleges and universities. Nevertheless, having many testing companies does not imply a similar diversity of college selection tests. The United States has 16,000 school districts and 50 states and hundreds of testing companies, yet for the most part, it relies on only one or two agencies for the design and administration of college selection tests. By contrast in Sweden performance ratings are designed and calculated by each classroom teacher largely independent of central controls. The United Kingdom has more than a dozen examination boards, though the population being examined is only 10% that of the United States. Australia has different selection examination systems for

each state, raising problems of inter-state comparibility. Japan has two sets of examinations. The first is uniform throughout the nation; the second is diversified, with different formats, emphases, different subjects administered independently by each college and each university on different dates, at different places, and at different times. In consequence Japan has more than a hundred different selection examinations, which requires a well educated, highly motivated and sophisticated consumer public.

The lesson one draws from the experience of OECD countries in the field of selection testing is that there is no single model. *There is no examination system whose design has not been influenced by the political culture in which it is situated.*

Aptitude or academic achievement tests?*

As genuine system-wide alternatives, the debate between academic achievement and aptitude tests is an American parochial issue. Scholastic Aptitude Tests (SAT's) were used in Japan in the 1950s and currently by several state testing authorities in Australia. SATs have not been used for admission into universities in Britain, France, Germany or Sweden. Despite the limited use of aptitude tests, the debate itself is useful for clarifying selection strategy in developing countries.

The SAT has four principal problems and two principal virtues. Though designed to measure basic aptitudes, the first drawback is that the SAT is, in fact, subject to coaching. Disagreement exists over the degree of the 'coaching' effect but not whether is exists.[1] The second drawback is the SAT's predictive ability. Japanese, Australian and American experience is similar in that university performance is more closely predicted by academic achievement than by academic aptitude.[2]

Third, the SAT creates a distance between what is taught in a classroom and the test. It limits the ability of teachers to prepare students and thereby lowers the 'feedback' effect of past test results. When performance is poor in a mathematics examination based on the curriculum, strategies of amelioration are significantly more evident and more clear than if the poor performance had occurred on the mathematics section of an SAT.

The last problem is that of measuring diligence. Performance on an academic achievement test is in part a reflection of a student's diligence, the ability to study hard for long periods of time, maturity of purpose. Diligence is one of the principle ingredients of university and later professional success. Diligence is not a trait easily testable by an SAT.

Nevertheless the SAT does have several characteristics to recommend its use in developing countries. It is neither criterion-referenced, nor based upon a single concept of subject excellence, and could therefore allow local school authorities to experiment with curriculum.

Its equity effects may also be a virtue. In developing countries the variability of school inputs is significant. Schools in the northeast of Brazil, for instance, have unit expenditures of about a sixth those in the southern Brazilian states. Moreover, the level of school inputs is more closely associated with learning in developing countries than in industrial countries (Heyneman and Loxley, 1983). Where school inputs vary widely such as in Nigeria, China, Indonesia and Brazil, academic achievement tests are likely to measure the opportunity to learn more than the ability to learn. Nations concerned about picking their future talent must consider the possibility that an aptitude test, such as the SAT, may be more able to overcome the local differences in school quality.

School-based or external assessment?

Class grades and school-designed achievement tests are very popular with teachers because they give them a professional sense of 'selection efficacy'. School-based assessments are on the other extreme from aptitude tests, for classroom activity is very close to what are thought to be appropriate criteria for selection.

It is argued in Sweden that school-based assessments are better able to measure what standardized tests cannot—student character, diligence, the will to succeed, and the change in these characteristics over time. It is argued that teachers know if a student is working up to potential; teachers know if someone with difficulty concentrating, has finally been motivated. For these reasons such assessments are

*Aptitude measures the ability to learn; achievement measures what has been learned.

excellent predictors of future performance and are growing in popularity in many OECD countries. School-based assessments have all but replaced nation-wide selection examinations in Sweden. In the U.K., school-based assessments are soon to become a requirement in university selection. By the year 1990 every British university will be asked to weigh the selection decision according to both the selection examination and the school-based assessments provided by the Local Educational Authorities (LEAs). Most American universities base their selection decisions upon a combination of school grades and SAT scores. In Australia and Japan, school-based assessment records are kept and are made available to the universities for making the selection decision. However the weight placed on such information differs by state and by university.

Testing experts feel ambivalent about school-based assessments. Such assessments are not 'standardized'. Criteria vary from school to school and from teacher to teacher. An 'A' grade, may constitute only a 'B' grade elsewhere. In some schools equal weight is placed upon physical education, fine arts and physics. In other schools only academic subjects figure into the grade point average. Moreover, school-based assessment results vary over time. In the 1960s student privileges and rights were in the ascendency in the U.S., Canada and Australia; curricular requirements were diminishing, and teachers tended to assess students according to less rigorous criteria than ten years earlier. This grade inflation, being idiosyncratic to some secondary schools and not others, made the selection decision more difficult. Colleges and universities had to adjust their treatment of school grades according to what they considered to be the reputation of specific secondary schools, or school systems. This presented considerable problems and numerous misunderstandings.

On the other hand there are efforts to minimize these problems. School-based assessments seem to function in Sweden because the student population is very small, thereby making school-based assessment logistically feasible; and there have been elaborate efforts to minimize the variation of criteria used by teachers to assign grades. Individual student results are compared to national 'norms' in each subject. Criteria for subject assessments are regulated by means of national 'guidelines'; and when extreme variance occurs, that is, when results stand out in some extraordinary fashion, a committee is ready to investigate the reasons and to recommend adjustments.

In large countries, the use of school-based assessments in lieu of standardized examinations would create severe problems. The logistics of grade recording would be more cumbersome. Since schools have vested interests in seeing that graduates continue with higher education, inflation of grades would be common. Questions would create political difficulties.

However, personal assessments of pupils can have some role in the selection decision.[3] They do provide more personal information. *Universities in developing countries ought to have personal information available for consideration during selection.*. This should include, in addition to grades, student efforts to support community projects; ability in sports, music and art. Each of these qualities help to make up the character of people whom universities may wish to select. Having available personal pupil assessments would make whatever special emphasis they choose, possible. This is particularly important now that university competition for students in specific specializations (and for income from student fees) is becoming more accepted and more common.

Multiple choice or non-multiple choice test format?

Multiple-choice questions have been in use since World War I and are dictated by virtually irrefutable necessities: they are easy to score and they can be given to large numbers of individuals quickly and cheaply. But 70 years after their first appearance, controversy remains. Many experts feel that the shortcomings of multiple-choice formats outweigh the benefits. The Cambridge University Examination Syndicate, for example, relies very heavily upon non-multiple-choice formats of examination. At both the 'O level' and 'A level' the Syndicate administers essay-format tests. Swedish classroom teachers, many Japanese universities, and some Australian states rely on essay formats, as do testing officials in France and Germany.[4]

If someone else grades the results, teachers generally prefer non-multiple-choice test formats. Essay questions, they feel, are more typical of normal classroom discourse and

represent a natural student-created response. Whether in school or work settings people are rarely faced with preprogrammed choices.

Multiple-choice formats appear to be more efficient at measuring skills at lower levels of the cognitive skill hierarchy; non-multiple-choice formats better (but not necessarily more efficient) at measuring skills at the upper end of the hierarchy, especially the skills of synthesis.[5]

According to their proponents, non-multiple-choice formats are amenable to the same types of reliability and validity checks as are multiple-choice formats. For non-multiple-choice formats to be standardized and objective, what is required is a carefully constructed system of professional test markers —as exists in Great Britain—with strong quality control, clear criteria of excellence and effective internal procedures.

Most experts feel that good multiple-choice test items are more difficult to construct. Such items have to be clearly written; their choices have to be plausible and comprehensive. Test designers need to know the principle on which the correct answer is based, but also the variant principles on which all conceivable incorrect answers are based. This requires considerably more research capacity than is normally available at the school level. Consequently, good multiple-choice tests are rare, while poor ones are common. Each of these arguments points towards the use of non-multiple-choice formats.

On the other hand, proponents of multiple-choice formats are quick to point out their strengths. Carefully designed, multiple-choice is useful even for testing higher-order skills of analysis, synthesis and evaluation. They can even be designed to assess skills involving creativity.[6]

Multiple-choice formats are amenable to scientific techniques of pretesting on an item-for-item basis. The Scholastic Aptitude Test (SAT) in the United States contains 20% 'dummy questions', questions placed in the test solely for purposes of experimentation. Thus when a test question is ready to appear in final form and is ready to be used in estimating ability, test designers have a fairly exact understanding of how that test item will perform even *before* the tests are actually administered. The degree of test bias is known ahead of time, for instance—against southerners, northerners, children who went to Catholic Schools, non-English speakers, or females. All of these can be predicted ahead of time with multiple-choice-test formats, carefully applied. This does not mean that the validity of non-multiple-choice formats is questionable. It does mean that if validity is questionable, the problems will more likely be discovered *after* test administration rather than before.

Few test agencies rely upon single format exclusively. The Cambridge University Examination Syndicate is known principally for its non-multiple-choice formats but it also uses multiple-choice formats. The Educational Testing Service, while known for its use of multiple-choice tests, annually administers thousands of written examinations as well. This implies that there is no universal correct format.

The perceived advantages and disadvantages of different test formats cannot be separated from the specific requirements for test application. The multiple-choice aptitude application used in the United States has one advantage over the method of achievement essay test application used in the United Kingdom. This is the ability to generalize from year-to-year and from subject to subject. Because examinations in Britain are designed by specialists against the criterion of subject matter excellence, the proportion of students who receive a 'first', 'second' or 'third' class pass, may vary from one year to the next. Variation may occur because students perform better from year to year, or because the difficulty of the examination may vary from year to year. By basing test results on scores that are equated over test administrations the American system is not plagued by the degree of variation in those scores.

However, the most important factor in determining the use of multiple-choice or non-multiple-choice formats is not test theory but test cost. Non-multiple-choice formats can be between two and five times more expensive to grade.[7] The larger the population taking the test, the larger the cost differences become. Though there are economies of scale in both formats, the marginal cost of adding a tested pupil in the U.S. is only a few pennies once the test is designed; the marginal cost in Britain is substantially higher due to the very high cost of test marking. For instance, Educational Test

ing Service in the U.S. annually examines many times the number of students examined by the Cambridge University Examination Syndicate. An essay format for every American university candidate would place the cost of the examination significantly above what it is currently. It is likely that these costs would have to be subsidized by the government if working class students were not to be excluded from university opportunity.

What does this imply for developing countries? China currently has a college entrance population at the level of the U.S., about 1.7 million. But while this represents 40% of the American 18–year-old cohort it represents less than 1% of the Chinese 18-year-old cohort. The proportion of the population going to universities in China is expected to rise, to 11% by the year 2000. This will entail a doubling, a tripling, and then a quadrupling of the Chinese examination candidates over the next 15 years. Higher education in many developing countries is in a similar growth take-off stage as in China. *The impact of these costs and the logistics of the testing process will require countries to move gradually away from an examination system based entirely upon non-multiple-choice test items, scored by hand, to an examination based principally upon multiple-choice items, scored electronically.*

Test questions: should they be public or private?

Countries have very different traditions on the question of test privacy, and these different traditions have had pronounced effects on the education system at large. In those countries where test items are made public, in Japan and the U.K. for example, the effect is to generate 'examinations students', students who spend a great deal of time studying portions of subjects likely to be examined and methods of response likely to elicit a good score. If the tests themselves are based upon spurious information, intensive preparation can detract from more creative study and generate a false standard in the utility of knowledge. In economic terms this constitutes an overinvestment (from the social point of view) on the basis of a bad 'market signal' (Klitgaard, 1986). Open publication of past questions exacerbates this market signal.

The SAT in the United States offers one example of test privacy.[8] Test booklets are numbered, returned after each test application, then destroyed. This allows test items to be used again with little danger that students have taken the test previously will be at a significant advantage, and it dramatically lowers the cost of examination design.

There are two problems with keeping tests private. If applied to academic achievement tests, the 'feedback' mechanism will not be as effective.[9] Classroom teaching is more effective when actual test examples are used. Moreover, there are political problems associated with test privacy. Test privacy places a 'distance' between the testing agency and the general public. The public is expected to accept that the test was a 'good test' on the basis of faith or, alternatively, on the basis of statistical information few can interpret—reliability coefficients, Kuder–Richardson indices, factor loadings. Much simpler and much more direct is for the public to be able to read and debate the virtues and drawbacks of each question.

Faced with severe social and political volatility, trust in institutions assessing further educational opportunity is essential. *Thus despite the cost savings which can be realized from test privacy, the wise choice in developing countries will be to make tests open and public each year and to readily encourage public debate on test content.*

Admission decisions: who should maket them?

University education is expensive and student places scarce. University places are heavily or completely subsidized by governments; hence it is understandable why governments in developing countries would wish to make the selection decision.

Governmental mechanisms for making admission decisions vary. Most countries—China and Tanzania, for example—do not review each individual decision. Instead they establish such strict entry criteria that universities have little or no latitude for making exceptions.

Such is not the case in the U.K., the U.S., Japan or Australia. In each of these OECD countries, governments have little or no role in the selection decision. To be sure, testing agencies such as those in Japan are often supported by governmental resources; nevertheless, admission criteria and decisions on individual admission are made independently by each university.

That universities could make their own

selection is revolutionary in centrally-planned economies. For many decades these countries have allocated students to universities regardless of individual preference. One side effect has been to determine *ipso facto* a hierarchy in the quality of a nation's universities. University planners, on the other hand, have long argued that universities need to develop individually, to develop specializations, for example in engineering, literature or economics. But such individual university direction cannot occur without control over who will be the students.

To most government officials in developing countries it is clear that universities within the U.S. or within other OECD countries are not identical in either course offerings or prestige. It is also clear that there is open competition among universities, with rises and falls due to local financing, dynamic management and the like. They can see that it is particularly important for less well known, often rural, universities to feel that they are not locked into their current status. For such a competitive university environment to exist, most experts recognize that individual universities should select their own students.

Some ask whether universities would make selection decisions fairly? Would they not select on the basis of family privilege? Would they discriminate unfairly against ethnic or religious minorities? Governments have to concern themselves with such problems, but with regard to testing agencies, the role, at least in OECD countries, is clear: it is to test fairly and to leave the selection decision to the university.

What about the testing agency itself, should it be independent of government control? Examination specialists tend to hold that testing is a question of technical professionalism, and that the testing agency is at its best when it is independent of government financing and political control. Reports from Japan, the United States, the United Kingdom and Australia confirm this view. From senior educational managers in developing countries there are caveats however. If universities control the admission decision, it assumes that universities will be independent in other areas as well—the content of study, the criteria for academic excellence, the direction of academic research. Who makes admission decisions cannot easily be separated from other, and wider, implications. Yet testing agencies in OECD countries are independent from government because they can respond to the university market directly. *If ministries of government totally control university programs there does not appear to be much point in creating an independent testing agency.*

GENERAL RECOMMENDATIONS

Professionalization of testing

Despite differences in size and financial resources, school systems in developing countries, without exception, require a professional capacity in the field of standardized testing and examinations. This has three prerequisites. The first is *research capacity*. Countries which set their own examinations must therefore develop them. Test development can be done in a haphazard manner, but the negative technical and political consequences are serious. Professional test development requires a permanent program of item design and experimentation; an on-going and permanent program of test result evaluation; an active item bank; an extensive set of professional committees in the higher education community as well as the teaching community in elementary and secondary education; and a minimum commitment to experimentation with new testing technologies and equipment.

Developing countries cannot afford to establish a research capacity in the field of examinations as an entity separate from their research capacity for more general purposes of testing. Yet usually these two functions are kept separate. The first is situated in a testing bureau, the second in a bureau of research and planning. Autonomy should not come at the expense of economy in scale. Computer and optical-scanning equipment need not be duplicated in examination agencies and educational research agencies. There is no reason to not jointly plan equipment acquisition or to jointly participate in technical training. To professionalize testing, developing countries must develop research capacity; but given resource constraints, *developing a research capacity in the field of examinations will not be feasible without coordinating its development with that of educational research in general.*

Second is *training*. The professionalization

of testing in developing countries will require a regular program of training, both internal and international. The required skills are often broader in nature than is commonly understood. Clearly a professional testing capacity requires psychometricians and statisticians. But it also requires survey research specialists, computer programmers, art and graphics designers, publishing specialists, production and distribution managers, cost accountants, social scientists. Each has a specific function in the production process. Eventually a product must be delivered at exactly the same time with no production errors, to hundreds of thousands of individuals without fail several times per year. The logistics of professionally managing a system of testing requires training.

Third is *equipment*. All modern systems of testing depend upon computing equipment. Even when manually graded by teachers, school-based assessments require a systematized memory. Essay tests require statistical records. 'Fill-the-blank' tests, creative design tests, oral tests, all require scores, statistical analyses, historical records, systematic reporting. On the other hand it is not necessarily true that the more sophisticated the equipment, the higher the level of testing professionalization. It is true to say that professionalized testing requires electronic equipment.

To justify the installation of equipment, however, requires an adequate source of development capital, a careful plan of acquisition and utilization, and a carefully-phased program to develop and to maintain the skills of technical maintenance. The degree to which the functions of examinations and standardized testing can be localized will determine the complexity of the policy toward equipment acquisition. For example each of China's 29 provinces will require its own expertise and facility. This is also likely to be the case with each Brazilian, Indian and Indonesian state. In some countries, such as the United States, professionalized testing is found down to the school district level. This is because school districts design and finance their own curricula. The general principle is that the *geographical unit in charge of curriculum design, school finance, or school selection will require a professional testing capacity. That, in turn, will require the necessary equipment, training and research capacity.*

Test administration and management

Selection examinations play an important role in a nation's economic development and are therefore to be considered an important national resource. This resource needs to be protected. The means by which this is accomplished in many OECD countries is to establish the examination agency as an administrative body, autonomous from government control, and from government finance. The U.K., the U.S., Japan and Australia allow the testing agencies to collect user fees for examinations. Income from these fees remains within the agencies themselves. the fees are small enough so that they do not inhibit educational opportunity, yet large enough for the testing agency to build its own research capacity and to set it own standards of technical excellence. Professional standards cannot be maintained if administrative budgets are subject to the ebbs and flows of ministerial politics and national economic exigencies. Income derived from testing is inevitably secure since it is based upon a guaranteed demand. *An independence of fiscal resources is a prerequisite for an independence of professional standards.*

Japan, the U.S., the U.K., Sweden and Australia have widely divergent examination systems, but there is no instance of a single examination agency with a national monopoly. The National Center for University External Examinations in Japan serves as the sole source for examining the first stage of university entrance, but the second stage consists of a plethora of different examinations and competitive examination bodies. The Educational Testing Service in the U.S., while widely utilized, is not the sole examining body. In each case the testing agency is forced to maintain professional standards not only with the general public but in an actual competitive sense, in relation with other testing agencies. *Competition among test agencies is healthy.* It generates informed discussion and it provides 'the consumer' (universities, schools and school districts) with alternative choices. Large countries such as Nigeria and China might be wise to develop a diversity of examination systems and agencies in some combination of local and national authority. Smaller countries may not be able to support such a diversity. In these instances consideration might be given to external institutions such as now exist on a

regional basis in West Africa and the Caribbean. If there is a lesson from the examination experience of the U.S., Japan, the U.K. and Australia however, it is that competition works for the common benefit.

Uses of standardized tests for system-wide assessment

The utility of testing goes much further than selection for higher education. To be sure selection examinations are politically the most well known of such functions and are certainly of high interest among educational managers in developing countries. Yet there are uses of testing for functions other than selection. These next three recommendations concern them.

(i) *Countries should measure achievement over time.* School systems, in countries large and small, need to record their progress—not only in the number of students educated, but in the knowledge students have acquired. Such a record cannot be attained solely by studying the results of selection examinations, for only the best students who have finished secondary education are eligible to sit for them. Such a record cannot be obtained solely by studying the results of school-based assessments, for standards may vary from year-to-year or from school-to-school. Measuring educational achievement over time requires a specially-designed test, with 'anchor' items to circumvent changes in curricula; and it requires that it be given to a representative sample of students at regular intervals.

The virtue of a 'national assessment' given on a regular basis is that it provides a relatively unambiguous benchmark on a nation's progress in teaching basic academic skills. A national assessment is particularly important for countries whose student population is on the rise, where enrollments are on the increase. In much of Africa, Latin America and Asia, for example, it is common to assume that students are learning as much as they once did when the system was smaller. The assumption is often false. But how false? In which subjects? In which types of schools? In which parts of the country? These types of 'monitoring' questions are essential for the managers of any educational system who wish to be informed.

(ii) *Countries will find it useful to compare the achievement of their students with the achievement of students in other countries.* Most OECD countries now recognize that their economies are interdependent and, in many ways competitive. Comparative statistics on trade, manufacturing, welfare, crime, public spending, private investment are considered a necessity. These statistics are studied carefully by managers both in and outside of government.

These demands for up-to-date comparative information are no less true in education. OECD countries share statistics on enrollments, progression rates, expenditures and, increasingly, academic achievement. Elementary and secondary school achievement is measured through commonly-designed standardized tests of science, mathematics and reading comprehension.

Using standardized tests to measure academic achievement across countries raises many serious dilemmas. On the one hand education is a local enterprise. One purpose of schooling is directed to preserving local culture, history, civic pride and language. At the same time, educational systems are attempting to teach many common skills. The ability to manipulate figures and interpret the written word are common goals of curricula in widely divergent countries. This pertains to wealthy and impoverished countries alike, to countries such as Sweden which can afford to spend US300/pupil each year on classroom reading materials and supplies, as well as to countries such as Bolivia which can afford to spend only 1% of what Sweden spends on each of its students.

Why should a developing country engage in research on academic achievement along with a country which can spend 100 times more per student? Aren't drastically lower levels of achievement inevitable?

The answer is 'yes' with respect to mean achievement but 'no' with respect to school effectiveness. International studies of academic achievement have taught educational managers an important lesson. Mean achievement comparisons across countries for science or mathematics are essentially meaningless unless they are weighted by two factors, the level of monetary resources brought to bear on the classroom, and the percentage of the age cohort enrolled. Although average mathematics scores at the 12th-grade level in Germany are higher than in the United States,

the United States is educating a higher proportion of its 18 year old citizens at the 12-grade level. Similarly although 14 year old students in Thailand have lower reading scores than 14 year olds in Japan, Japan is able to invest ten times the level of monetary resources per child in reading materials. Interpretations of school achievement are not on the basis of 'Olympic records', but rather in relation to to the constraints operating on the system.[10]

Results have been revealing. Some very 'effective' school systems—judged by these standards of resource handicaps—have been found in developing countries, particularly in Asia. Regardless of the specific results however, the main point is the managerial function which such cross-national testing serves. It can point out the strengths and weaknesses of an educational system in relation to other systems. It can inform the managers of educational systems whether, by comparison to others, they are teaching certain skills too early or too late;[11] whether certain classroom equipment or pedagogical philosophies are effective; whether certain managerial innovations—cross-age teaching, centrälized learning centers, modular instruction—are functional. Countries should monitor the process of innovation in the field of education like they monitor changes in industry or other parts of their economies.

Using examinations to improve classroom pedagogy

It is true that teachers teach to an examination. National officials have three choices with regard to this 'backwash effect': they can fight it, ignore it, or use it. *The recommendation to developing countries is that they design examinations with well balanced pedagogical principles in mind and that they maximise the influence of examinations on classroom teaching.*

For many years it has been popular among educators to fight the use of examinations on the assumption that the effect was harmful to good teaching. The reason other selection criteria have been used to replace examinations was because rote memorization was assumed to be the dominant characteristic of the 'backwash' effect.

Examination agencies are those most guilty of ignoring the 'backwash' effects of examinations. The management is often made up of examination technical specialists rather than educational management specialists. Examination specialists see their role as external to the education system, as auditors. They do not see the quality of classroom teaching as their responsibility.

However, examinations can be used to promote positive pedagogical ends. In these instances educational managers allocate resources to provide examination information to schools in detail before and after the test; and they make certain that the skills needed on an examination are wider than those of rote memorization. One such instance is the example of Kenya in which considerable attention is given by the Ministry of Education to explaining the thought patterns behind wrong answers (Somerset, 1987). 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 education.

Test content and format

Multiple-choice. In Nigeria, Indonesia, India, Brazil, China and in other large and populous countries, the exigencies of economy and logistics will determine test format. For instance, as experience is gained and as the computers and optical scanning equipment are acquired at national and provincial levels, these countries might utilize multiple-choice formats extensively. The key is to introduce multiple-choice formats only after the requisite technical experience has been gained.

Multiple-choice test formats are not always a sign of good testing; therefore a multiple-choice test is not a sign of 'modernity'. Developing countries need not relinquish what common sense dictates: the written essay must remain. But the requirements for standardization are substantial. They are particularly complicated when large-scale testing is being conducted in languages which have not long been utilized for such purposes—Kiswahili, Quecha, Indonesian, Filipino, Nepali, Creole. All languages have local dialects and terminological variations; all languages have stylistic alternatives. But in those languages which have long been used in large-scale testing—Japanese, French, and English are examples—tradi-

tions of standardization have been tested through time; they are widely acknowledged and widely understood in the school system. Attempting to judge excellence in an essay when the tradition is only now being standardized, is an endeavor which involves high levels of risk. With a multiple-choice test item, the strengths and weaknesses are visible for all to see. With an essay test item the difficulties are more hidden and begin to emerge only after the item is set and the test taken. *Developing countries which employ a national language which has not been used extensively for educational testing may be wise to shift to a multiple-choice format as rapidly as possible.*

Scholastic Aptitude Testing in developing countries. Only one OECD country reports the regular use of the SAT, and that is the United States. Its use in the United States, however, is largely determined by the style of educational governance rather than the virtues of aptitude testing. There are 16,000 different school districts in the United States, each feeling strongly about its jurisdiction over curriculum. Since no agency can successfully market 16,000 different versions of an achievement test, these authorities therefore must settle for a non-curriculum-based alternative. All debate — pro and con — about an ability test vs a curriculum-based achievement test must be discounted by the fact that the origin of the debate is basically confined to the U.S. which because of its own political circumstance, may have no feasible alternative.

By all admission, the SAT is difficult to design. It requires a strong research base and a large supply of the technical skills. Reasons of cost require that past tests be kept private. Because of its privacy and non-applicability to specific curricula, its feedback influence on the classroom, by comparison, is small.

Nevertheless in conjunction with curriculum-specific grades (or achievement tests) the SAT is a reliable predictor of future academic performance. It is also able to identify academic talent in impoverished schools where the opportunity to learn has been minimal. Large countries, such as China, might experiment with the SAT on a regular basis. In a decade, the local research base required to support the SAT may be available and would not have to be imported. Moreover, the statistical data generated from SAT experiments often make good cross-checks on the validity of achievement tests currently in use. Regular statistics on the predictive validity of the local SAT vs the national achievement test can be very useful information. Nevertheless *it is unlikely that the SAT would be of use in developing countries as a standardized mechanism of selection for higher education.*

Skills to be tested

There is considerable agreement that academic skills are not unidimensional, and that skill categories are hierarchical. But what is an appropriate distribution of test items across this hierarchy? What proportion of a selection examination should be dedicated to testing the skills of knowledge recall? What proportion should test application skills? What proportion should test the skills of synthesis?

In China, the government English language test distributes items according to a normal distribution, with 15% of the questions testing knowledge recall at one end of the hierarchy, and 12.5% of the questions testing the skills of synthesis at the other end of the hierarchy (Table 1).

Table 1. China: English proficiency test — test items by type, 1985

	%
Knowledge	15
Comprehension	26
Application	25
Analysis	21.5
Synthesis	12.5

In Kenya the categories have shifted over time. In 1973, 74% of the question items on the primary leaving examination tested the skills of descriptive recall. By 1976 the proportion had fallen to 23%; the difference was made up by proportional increases in the items devoted to higher-level reasoning (Table 2).

Table 2. Kenya primary leaving examination: test times by type and year

	1973 %	1976 %
Descriptive	74	23
Explanatory	18	28
Observation	8	21
Reasoning	0	28

There is no consensus on the distribution of items across skill hierarchies. Nevertheless, *developing countries should make certain that all categories are tested and that the proprotion devoted to each category is well known to all teachers prior to the test.*

SUMMARY

It is fair to say that the policies on examinations in developing countries have changed since the 1960s, from resisting their effects to using their effects. Nevertheless many problems remain. This paper has attempted to summarize some of the most common issues and to make recommendations. To reiterate:

1. *General conclusion*

No system of examinations is designed on technical grounds alone, each exists in a political environment. The system of aptitude testing in the United States exists because of the complex political prerogatives for communities to control their own curriculum. School-based assessments can function in Sweden because of the modest logistical prerequisites and the consensus on criteria achievable in a small monoculture. Non-multiple-choice formats can exist in Britain because the number of test-takers remains manageable and the definition of academic excellence has but modest variation from one university to another. Multiple testing by individual colleges can function in Japan because of the level of sophistication and motivation of the test-taking population. What this implies is that there is no OECD model which can be transferred without forethought and adaptation. The necessity for forethought and adaptation is particularly important for ex-colonial territories in the French and British Commonwealth.

2. *Universities*

The mechanism of selection will affect the quality of universities and therefore a nation's future. If universities are expected to increase levels of self-financing (which they are) and to be competitive with each other, or perhaps internationally (as they are), then universities should be given the responsibility of selecting their own students. Moreover, they should have access to school-based records of student accomplishments to assist them in their choice.

3. *Test agencies*

Testing is a profession, but it is highly susceptible to political interference. To a large extent, the quality of tests rests on the ability of a test agency to pursue professional ends autonomously. Agencies should therefore have their own source of finance from test fees. In larger countries competition among agencies would be healthy. On the other hand test agencies subsidized by the public sector should be expected to fulfill public functions. Among these functions should be to establish a strong system of analyzing test results and feeding this information back into the school system. Testing agencies should also share technical skills (item design, computing programming, etc.) and equipment with educational research functions of other agencies.

4. *Tests*

Where the test-taking population is high, geographically dispersed, culturally heterogeneous, or where the test employs a new national language, the test itself would benefit from a multiple-choice format. Pedagogical effectiveness would be maximised if questions were based upon curriculum; if they were open *ex post facto* to public scrutiny; and if they were to include all levels of skill hierarchies, from recall to synthesis.

5. *Countries*

Perhaps the most significant challenge of developing countries in the field of education is the management of the substantial investments already made in the system. Better management will require systematic use of three mechanisms: (i) feedback of examination performance; (ii) comparisons of academic performance over time; and (iii) comparisons with countries in other parts of the world.

NOTES

1. On the other hand, many proponents of academic achievement testing view 'coachability' not as a drawback, but rather as a virtue. One reason why the SAT was rejected in Japan, for instance, was because it was not as coachable as an achievement test.

2. On the other hand, ETS points out that university performance is more closely predicted by scores on the SAT in combination with academic (school-based) grades than by the SAT alone or by grades alone. Most American colleges and universities use a combination to make selection decisions.

3. School-based assessments are used in those cases where students are moving from secondary school to Swedish university; they are less influential among adults who wish to enter university.

4. An essay is only one non-multiple-choice format. Fill-the-blank, short answer, and 'design projects' are others. For example at the 'A' Level in highly selective fields such as engineering, the Cambridge University Syndicate administers tests of originality. This consists of a project on which each individual will work for up to six months.

5. In each testing situation there exists a hierarchy of skills which tests wish to measure. Most simple are (i) knowledge and vocabulary awareness skills. Others are: (ii) comprehension, (iii) application, (iv) analysis, (v) synthesis, and (vi) evaluation skills. Synthesis skills consist of those which one utilises to combine many pieces of knowledge, facts and principles, with the end result being a cogent single product.

6. If creativity is defined as 'novel ways of looking at a problem', multiple-choice-test formats can be used. Responses can be thought through ahead of time by the test designer. However, if creativity is defined as 'unique ways of looking at a problem', multiple-choice-formats would be inappropriate; in fact, impossible since the criterion of excellence is something not yet invented.

7. Costs are determined by the size of the test-taking population and the degree of prior research required for item development, among other factors.

8. 'Open Testing' legislation in the U.S. has made it possible for students to have access to their results, item by item. However, the procedure calls for the release of these results on the basis of individual demand. Unlike a public test, items which appear on an SAT one day are not normally published in the national newspapers the next.

9. Where test 'leaking' is a problem, such as India, keeping test questions private means that the rich and privileged would be at an additional advantage since they are usually the main market for such 'leaks'.

10. That one golfer has a score of 50 vs another golfer's score of 40 means very little unless the two are playing with the same handicap, on the same course, under comparable conditions. On the other hand, knowing only that a golfer has a score of 50 is meaningless unless it is compared to something else. Many school systems in developing countries don't even have a 'score'; they have no measure of what has been learned. But among even those with a score, most have nothing by which to compare it.

11. OECD countries have discovered that Chinese children are expected to acquire certain arithmetic functions in grade two, which, in accordance with certain Western theories of child development, should not be taught until grade four.

BACKGROUND PAPERS*

Testing for selection to university

1. Hidano, T., Admission to higher education in Japan.
2. Reddaway, J., Examinations for university selection in England.
3. Solomon, R., Admission to higher education in U.S.: the role of the educational testing service (ETS).
4. Keeves, J., Public examinations in Australia.
5. Marklund, S., Education in Sweden: assessment of student achievement and selection for higher education.
6. Lu Zhen, A brief introduction to the system of higher school enrollment examinations in China.

Testing for the improvement of educational management

7. Lapointe, A., Assessing the quality of education over time: the role of the national assessment of educational progress.
8. Keeves, J., Cross-national comparisons in educational achievement: the role of the international association for the evaluation of educational achievement (IEA).
9. Somerset, A., Examinations as an instrument to improve pedagogy.

REFERENCES

Heyneman, S. P. (1980) Investment in Indian education: uneconomic? *World Development* **8**, 145–163.

Heyneman, S. P. (1986) The search for school effects in developing countries. E.D.I. Seminar Paper No. 33.

Heyneman, S. and Loxley, W. (1983) The effect of primary-school quality on academic achievement across 29 high and low-income countries. *American Journal of Sociology* **88**, 1162–1194.

Heyneman, S. P. and White, D. S. (1986) *The Quality of Education and Economic Development.* World Bank.

Kirmani, N. *et al.* (1984) Effects of increased market access on exports of developing countries. IMF Staff Papers 31, 4.

Klitgaard, R. (1986) *Elitism and Meritocracy in Developing Countries.* Johns Hopkins University Press, Baltimore, MD.

Pinera, S. and Selowsky, M. (1981) The optimal ability–education mix and the misallocation of resources within education. *Journal of Development Economics* **8**.

Somerset, H. C. A. (1987) Examination Reform in Kenya. World Bank Education and Training Series No. EDT 64.

*Each paper is to appear in Heyneman, S. P. and Fagerlind, I. (eds) (forthcoming) *Improving University Section, Educational Research and Education Sector Management in Developing Countries: The Role of Examinations and Standardized Testing.* World Bank, Washington, DC.