DEVELOPMENTS IN ITEM BANKING

BRUCE CHOPPIN

Paper given at the first European Contact Workshop held at Windsor, UK, in June 1976

Published in "Monitoring National Standards of Attainment in Schools", R. Sumner, Ed., Slough, UK: NFER

Abstract - Item banks can be used to develop effective and efficient systems of tests and examinations for the purposes of assessing the achievement of individual students, of monitoring changes in the curriculum and for evaluating other educational innovations. Full exploitation of the advantages inherent in the item bank concept depends on the adoption of an explicit model of test-taking behaviour, such as that proposed by Rasch. Three diverse applications of item-banking, based on this measurement model, are presented.

WHAT ARE ITEM BANKS?

The term **item bank** should be understood to mean a collection of test items organised and catalogued in a similar way to books in a library, but also with calibrated data on their measurement characteristics. This means that when a test is constructed from a sub-set of items taken from the bank, these calibrations can be used to determine the psychometric properties of the test. In addition scores from two tests made up of separate items from the bank can be interpreted one in terms of the other, since both are based on the same set of item calibrations. So far most of the item banks that have been constructed have been of the multiple-choice type, and most have been concerned with the measurement of school attainment, although the item banking concept need not be restricted in either of these ways. Item banks provide the test constructor, whether he be an individual teacher or a member of a National Examination Board, with access to a very wide range of items. From these he may select any one of an astronomical number of alternative groups to use as a test with specifiable psychometric characteristics. Since the basis of item calibration is common to all the possible groups, the scores produced may be translated to a common psychometric scale and hence interpreted almost as though the tests were parallel.

It should be noted that the above definition of item banks would exclude mere collections of items which have been assembled as an aid to the sharing of creative

ideas between examiners in different institutions. The term 'item pool' would be used for any such collection if it lacked the necessary psychometric calibration to permit the use described in the foregoing paragraph.

The chief virtue of a complete item banking system is its flexibility. In theory at least it enables people who wish to make measures of achievement to have access to a wide range of well documented testing materials to cover a whole variety of situations. Though it is in many ways rather more complicated to operate than would be a set of standardised tests, it offers several advantages. When one has a large pool of test items upon which to draw, test security is not the same problem as it sometimes is in a standardised testing situation. Furthermore, teachers can 'design' their own test and yet have the results readily related back to some larger reference framework. This should improve the quality of classroom testing and also help teachers to appreciate the value to them of sound educational measurement.

In fact I see applications of item banking in three separate fields; fields which are certainly not mutually exclusive. Firstly I see them being extremely valuable to teachers who want to design their own high quality assessment instruments, but who do not have the time or the skill to develop an achievement test from the beginning. Many teachers could undoubtedly make good use of a well organised item bank, and because of the time it would save one might hope that not only the testing but also the teaching would improve.

Secondly, item banks may prove to be especially well adapted for meeting the needs of criterion referenced evaluation as mastery learning strategies become more and more widely used in education. When item results rather than test scores become the focus, then an item-based measurement resource would seem to be the natural answer.

Thirdly, more than half the countries of the world are only now beginning to develop large school systems which can offer something like universal education to their young citizens. In these countries where financial resources and psychometric expertise are both in very short supply the need is for cheap but comprehensive systems of educational assessment, and I believe that item banks can provide this. An area that is still to be explored is the extent to which item banks may be effectively shared by different countries and different educational systems, but there seems good reason to suppose that this can be done. Such a facility would be important in regions such as Latin America or East Africa where language barriers are not insuperable.

Within Europe the language question cannot be ignored, and it is as yet too early to say whether test materials may be freely translated from one language to another

without their psychometric properties being substantially altered.

More has been written about the potential advantages of item banking systems than about the practical results of introducing them. Wood and Skurnik (1969) in their influential book on item banking describe in detail how it could serve the British desire to have both national and school-based assessment of secondary school pupils. Scriven (1967), in an article which has had a major impact on the theory and practise of curriculum evaluation, explains how item banks could completely change procedures for both formative and summative evaluation, providing, at the same time, for continuous assessment of student progress. The difficulty in implementing these ideas until very recently has been the lack of an adequate methodology for handling measurements based on responses to individual items. This paper describes such a methodology.

ORGANISATION AND CALIBRATION OF AN ITEM BANK

Just how should an item bank be organised and what data should be stored together with the items. It is usual of course, even in otherwise unstructured item pools to record some information as to what each item is supposed to measure. Here perhaps we can distinguish two quite different types of item banks. In one, each item essentially measures some different aspect of achievement. Each is perhaps concerned with some criterion task, performance of which would demonstrate mastery of a particular objective. Such items are intended to be used and interpreted individually. The other sort of bank will contain substantial numbers of items which purport to measure the same dimension and this dimension is probably quite generally defined as being for instance 'Achievement in Geometry' or 'Knowledge of German Vocabulary' or 'Understanding Scientific Principles'. In these circumstances it is expected that groups of items will be extracted from the bank and used to form an ad hoc test to provide more or less precise measurement for the trait in question. The bank must contain information as to what each item is supposed to measure, but this in itself is not enough. If results on an item are to be interpreted then one needs to know how difficult the item is and to what extent it discriminates between people of different ability. Unfortunately, conventional measures of item difficulty and discrimination are 'sample-bound', which is to say. they are extremely dependent on the nature of the sample of people who provide the data. One way round this problem is to try out test items on a wide variety of different types of people, and to store the results for each group of people together with the description of each group for future use. This approach has been used in the past for a wide variety of standardised tests, but it is basically not very satisfactory. There are problems in adequately defining the groups of

people for whom norms are calculated and also problems in identifying a particular testee as a member of a particular group.

Gulliksen (1950) hinted at a different approach to the problem:

> "A significant contribution to item analysis theory would be the discovery of item parameters that remained relatively stable as the item analysis group changed; or the discovery of a law relating the changes in item parameters to changes in the group."

Much recent psychometric research has concentrated on the identification of item parameters with this property.

An important characteristic of an item bank is that, just as with a money bank, different units deposited and withdrawn from the bank may be related to one another by means of a well-defined currency. Normal methods of test analysis do not provide this property. A test score is normally interpretable only in terms of the particular content of the test. Scores on two different tests cannot normally be compared directly, but only through the device of transforming to percentiles of the normal distribution or something similar - and this results in the interpretation of one individual's score being dependent on who is included in the sample for the norming of the test.

Such sample dependence is very limiting for measurement. It would be intolerable if to compare the lengths of various pieces of string one had to measure all of them with the same ruler. Intolerable also if the length we found for a particular piece of string was dependent upon which other pieces of string had been measured. This, however, is the situation we face with traditional forms of test analysis.

In the search for 'sample-free' test parameters it was soon realised that any model for item response would need to be probablistic rather than deterministic if it was to adequately represent reality. Earlier work on test taking behaviour has shown conclusively that, without taking a vast number of other factors into account, a simple model will only be able to estimate the probability that a particular individual will correctly respond to a particular test item, and not definitely to predict whether or not he will succeed.

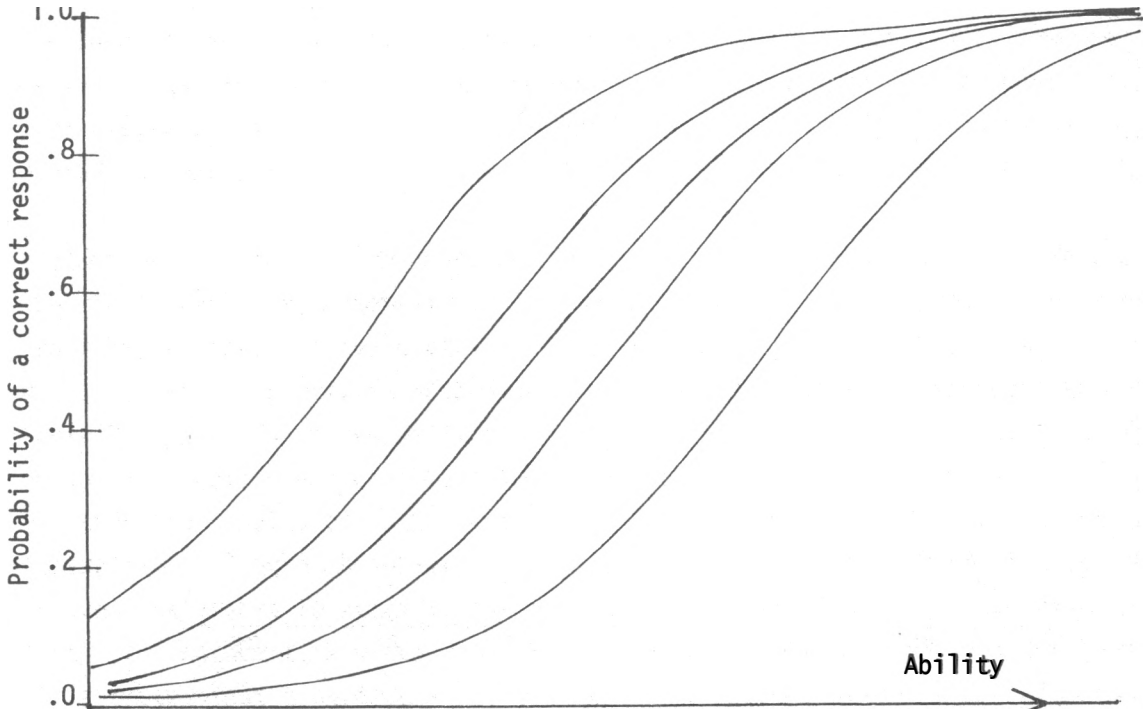
Of the various possible stochastic models for describing test taking behaviour, we will here consider only one family with particularly simple measurement characteristics. This is the one now being used in a number of different item banking projects, and it is the Rasch model described by Dr. Willmott in his paper.

The basic theme underlying this model is that 'difficulty' of an item should be defined in a special way. One result is that if you, or I, or anyone else, be confronted with two test items of different levels of difficulty then our probability of responding correctly to the one of greatest difficulty is always less than our probability of responding correctly to the other. Further, since difficulty is defined in terms dependent of whether you, or I, or anyone else, faced the items, the 'relative difficulty' of the two items is preserved. The first part seems no more than common sense, but the second is slightly more difficult. If item 'A' is easier than item 'B' for me then it must be so for you too and for everyone else. Similarly if you have a greater probability than I of responding correctly to item 'A' then you will also have a better chance at item 'B' or indeed any other item in the set. Any group of items scaled by the Rasch model can be ordered with regard to their difficulty independent of which people will be exposed to them, and conversely any group of people whose abilities are assessed with the Rasch model can be ordered according to their probability of responding directly to a particular item without regard to which item is used.

A convenient way of summarising the behaviour of a test item is the 'item characteristic curve' (ICC). This plots, for a single item, the probability of a person responding correctly to the item for persons of different ability. In general the probability is low (near zero) for people of low ability, and high (near one) for people of high ability. The slope of the curve at its central point gives some indication of the item's discriminating power. The order relationship of 'relative difficulty' within the Rasch scheme means that the ICC's for a set of items take on a quasi-parallel form as shown in Figure 1. The curves do not cross.

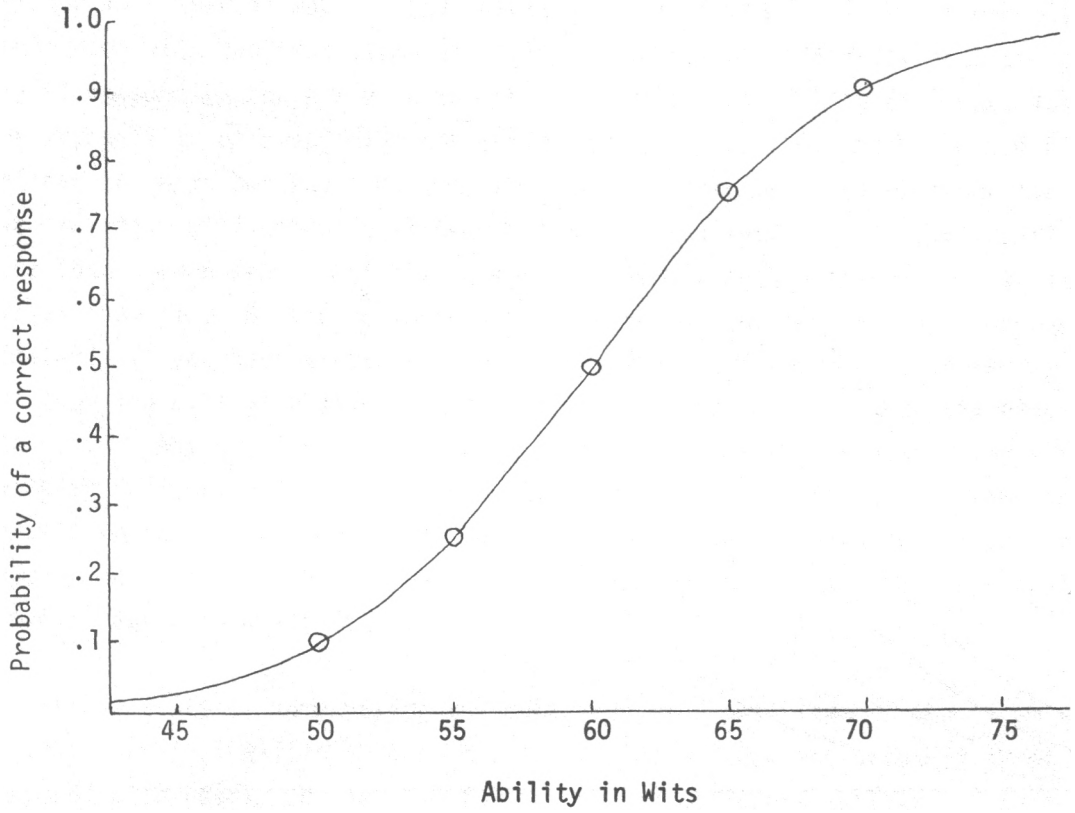
Now this is the theory behind the model and it does not exactly match what one finds in real life and with actual test data. Firstly, it clearly does not hold for groups of items measuring different subjects. If item 'A' concerns 'Euclidean Geometry', and item 'B' - 'knowledge of Spanish Vocabulary', then it is easy to identify types of people for which either one but not the other item is easy. In addition one cannot usually mix items which allow the answer to be guessed with those which do not, as this offends the relative difficulty rule for persons of different abilities.

FIGURE 1 - Item Characteristic Curves According to the Rasch Model



Nevertheless the model has proved to be robust enough to scale large numbers of test items in particular subject areas. It seeks to describe the outcome of a person confronting a set of test items in terms of parameters for the difficulties of each of the items involved and a parameter for the relevant ability of the person. Both ability and difficulty are measured in the same units - the 'wit' is frequently used. Biology test items are calibrated in biology wits; chemistry items in chemistry wits. The calibrated difficulty of an item is a stable property of the item, assumed to be unaffected by the nature of the people who attempt it. An individual will have a certain amount of ability in biology (measured in biology wits); and ability in chemistry (measured in chemistry wits) and so on. A person with ability 60 wits in biology would have a 50% chance of answering correctly a biology item of difficulty of 60 wits. For an item of 55 wits his chance of success would rise to 75%; for an item of 50 wits 90% and so on. The probability of success for various abilities attempting an item of 60 wits is shown by the item characteristics curve in Figure 2.

FIGURE 2 - The Item Characteristic Curve for an Item of Difficulty 60 Wits



In an item banking context it is necessary once a pool of scalable items have been constructed to perform a calibration; that is to estimate the parameters of difficulty for each item. For this a certain amount of trial data is required, but it should be noted that it is never necessary to have a sample of students attempt all the items.

Once a calibration of the items has been carried out, it is comparatively simple to calibrate the whole range of possible test scores which could be obtained on any sub-set of items withdrawn from the pool to make a test. This leads to straight-forward estimates of ability for individuals that have the desired characteristics mentioned earlier.

When item data is stored in the bank ready for use it would be necessary to include the difficulty calibration, along with information regarding the item content, directions for administration, for scoring etc. There is, in theory, no other psychometric data to be included, since topic and difficulty level are the only information needed to interpret test results. It is, however, rather too early to say whether or not such an idealistic approach will suffice in practice. Rasch-

scaled item banks are neither norm-referenced nor criterion-referenced, but fall neatly between the two concepts. For a strictly hierarchical subject or for measurement along a single clearly defined dimension, the distinction is not important. For more complex subjects (e.g. Mathematics treated as a whole) discrepancies from the model do appear whenever performance on particular objectives is compared. Research into this matter is continuing, and it appears that for diagnostic use a secondary analysis of test results to locate discrepancies and measure the residuals may be appropriate.

Another unresolved problem concerns the scope of the item bank. Should one have the single bank for mathematics, with one scaling scheme for items in arithmetic, geometry, algebra, analysis etc.? Does it make sense to talk of someone's mathematical ability, or must it be broken down into these sub-areas in order to be meaningful? Here there are still disagreements, but it is not difficult, for the time being, to organise and calibrate a mathematics item bank in two ways. If it turns out that the results of both are essentially the same then the more complex one (of separately scaling each sub-set of questions) can be dropped. As far as possible, it seems desirable to stick to the notion that mathematics items have fixed difficulty levels (in mathematics wits) whether they concern algebra, arithmetic or geometry. If it proves that a particular individual performs much worse on the algebra items than he does for example in geometry, it seems permissible to deduce that his mathematical ability is slanted away from algebra towards geometry. This may be a better way of proceeding than to have entirely different estimates of ability (in algebra wits and in geometry wits) that cannot be directly compared.

I shall now briefly describe three different applications of Rasch-scaled Item Banking. I have selected them not because they are the best in the field or even because they are generally representative. They are here because I feel I know each of them well enough to describe them to you and each illustrates a different type of application of the basic banking concept.

APPLICATION 1

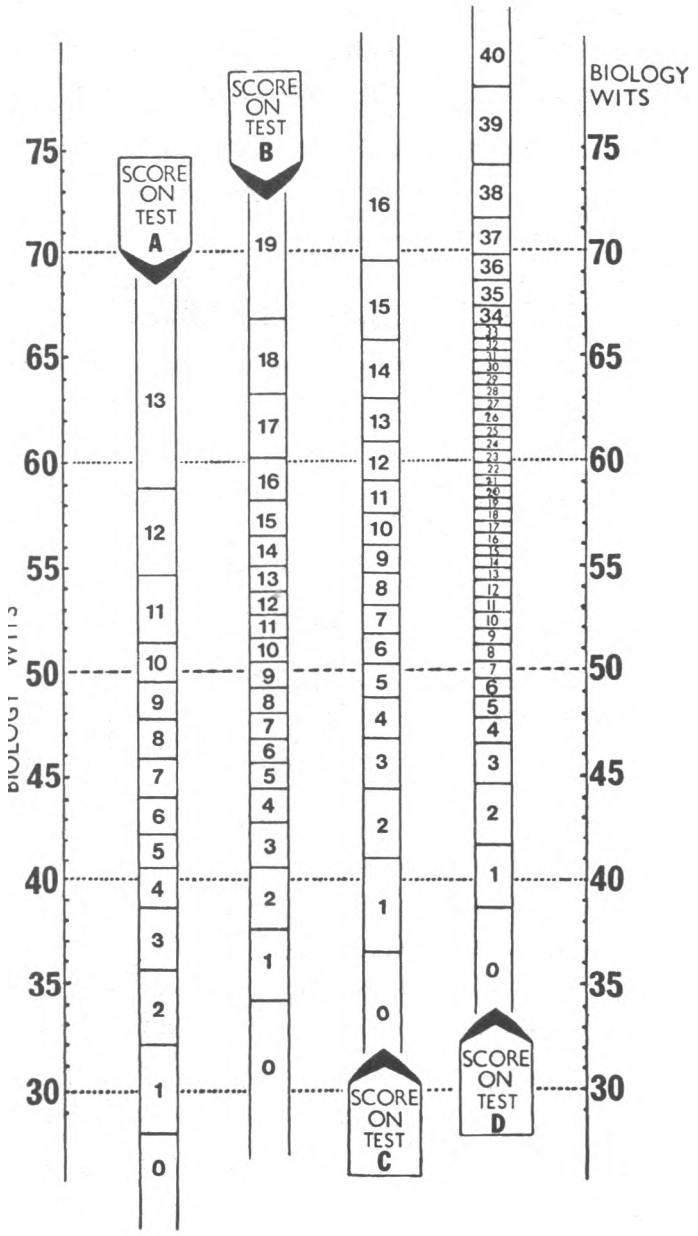
The first example is a piece of exploratory research. This piece of work used the science item bank created by the International Association for Educational Achievement (or IEA for short) for their cross-national study of achievement in science. The main report of the study appeared in 1973 (Comber and Keeves) but the work I shall now describe is an extension using children from the English school system.

For this application the IEA science item bank was divided into three main areas: Biology, Chemistry and Physics. Each of these contains items covering a wide range of ability with at one end some items suitable for children in primary schools and at the other some appropriate for those preparing to enter university. Four tests were constructed with items drawn from the bank. Each contained sub-tests in biology, chemistry and physics, but for simplicity I will only refer to the biology results here. Test A was intended for ten year old students who in England are mostly to be found in Grade 5. Test B was regarded as suitable for those in the early years of secondary education while Test C was intended for those in the twelfth grade. In addition, a particularly difficulty Test D was created to be administered to those pre-university students who had specialised in the study of biology. Each item in these tests had already been calibrated as members of the total item bank (biology sub-section), and this enabled calibration of the scores on each of these new tests in terms of ability.

Figure 3 shows the results of the set of calibrations. It gives equivalenced scales across the tests as an aid to the interpretation of the scores. They can be read in either of two ways. First, one may say that a student with 45 wits of ability in biology would be expected to attain a score of seven on Test A or a score of five on Test B, and so on. Alternatively, one can argue that a score of ten on Test A is evidence of an ability of 50 wits whereas a score of ten on Test C suggests an ability of 57 wits. A good deal of information about the tests can be discerned from Figure 3. Notice, for instance, that the tests are of varying length; Test A has 13 items, Test B has 19, Test C has 16 and Test D has 40. This does not cause problems for the analysis, and it is clear how the extra precision in measurement is obtained when the number of items is increased. Note also, how much more imprecise are the measurements made towards the ends. On a 13 item test such as Test A, although there are 14 possible scores, it is only possible to make 12 estimates of ability. For somebody who scored 13 (that is; all responses were correct), all we can say is that their ability is probably greater than 59 wits. Without additional information, it is not possible to put an upper limit to the ability estimate. Similarly, somebody who scored zero on Test A probably has an ability of less than 28 wits. How much less we do not know. Finally, note the very considerable degree of overlap that exists between these different scales even though, for example the items in Test B were substantially more difficult than the items in Test A.

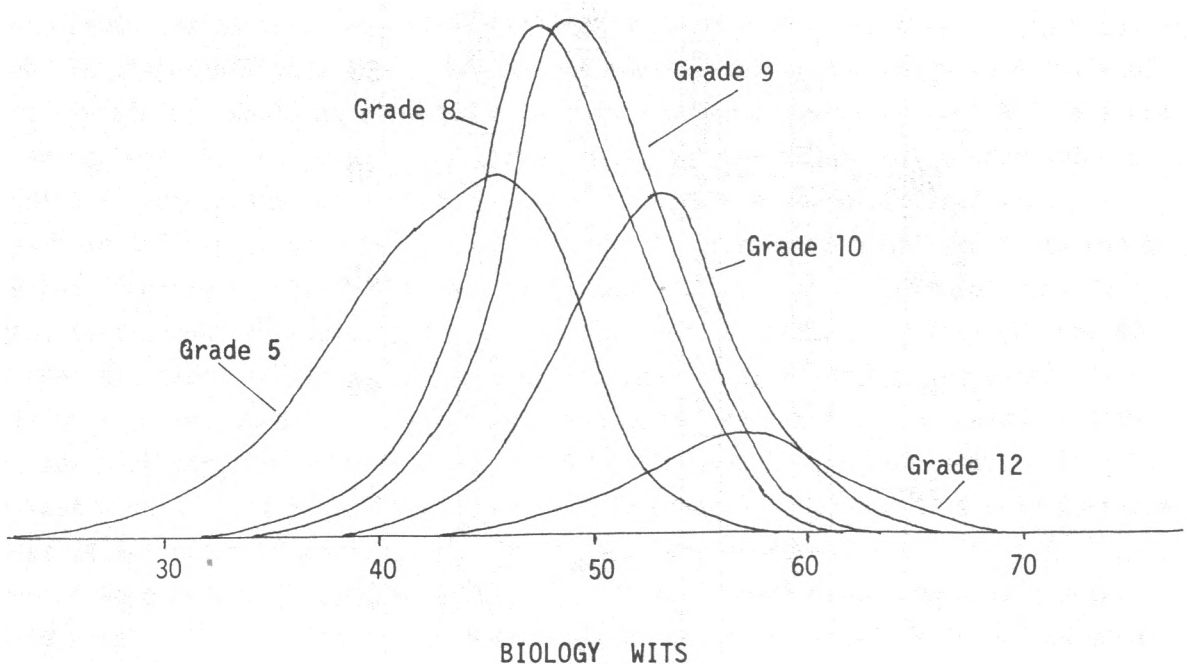
The appropriate tests were given to random samples of English school children aged ten (Test A), in Grades 8, 9 and 10 (Test B), and Grade 12 (Test C). Some members of the last sample who were specialising in biology were also given Test D.

FIGURE 3 - Calibration in Wits for Four IEA Biology Tests



Scores on the test were then converted into wits and frequency distributions for ability were calculated. They are shown in Figure 4. The areas under the curves for Grades 5, 8 and 9 are the same since virtually 100% of the children attend school for these Grade levels. However, at the time the data were collected, Grade 10 was not compulsory in England and some 30% of the 16 year old children were no longer at school. The smaller area under the curve for Grade 10 reflects this. Similarly for Grade 12 the proportion of children still in school was very much lower.

FIGURE 4 - Frequency Distributions for Ability in Biology



The information contained in these related distributions has not yet been fully digested. The spread of achievement in Grade 5 (which, for most children, is before any formal instruction in science begins) is clearly greater than that occurring in Grade 8 and 9, (when all the children have received about the same basic instruction in biology). By Grade 12, when specialisation and a plethora of elective courses lead again to great variation in the amount of exposure to biological ideas and thinking, the spread of achievement is again very high. Further, the growth in achievement from Grade 8 to 9 and 9 to 10 appears remarkedly consistent especially if one remembers it is, by and large, the less able part of Grade 10 that had dropped out of school before testing took place. The annual increase in achievement appears to be between one and two wits, and indeed the growth in achievement from Grade 5 to Grade 8 also appears to support this. The biologists with whom I have discussed these results so far find less than two wits a startingly small improvement for something more than 100 hours of classroom study. A typical student with a probability of 0.50 of responding correctly to a typical item at the beginning of the year, finds this probability increased to only 0.57 at the end.

Of course this is only an average figure. There will no doubt be some specific parts of the curriculum where mastery is achieved and the 0.5 might grow to 0.9 but this implies that there will be many more areas where growth is non-existent. The total amount of learning appears small.

APPLICATION 2

My next example concerns the use of item banks in curriculum evaluation along the lines proposed by Scriven in the article already cited. Some nine years ago a School Reform was initiated in Israel. This involved the gradual establishment throughout the country of a system of comprehensive middle schools to cater for Grades 7, 8 and 9. The old system of eight years elementary and four years high school was to be progressively replaced by a three tier system - six years, three years and three years. Although a major purpose of this reform was to bring about a greater degree of mixing of pupils from different social groups, the opportunity was also taken to modernise the curricula for the middle grades.

As part of the Goverment's plans for monitoring the effects of these changes it was decided to administer achievement tests to a large sample of pupils both in the middle schools and in a control sample of unreformed elementary schools as they completed each school year. This was repeated at the end of the 6th, 7th, 8th and 9th grades. The longitudinal study involved testing in a variety of school subj-

ects, but here I will only report to you some of the results for mathematics.

A large pool of test items for these grades was compiled: partly from existing tests and examinations, partly by the writers of the new curricula and partly by the team in charge of the monitoring. It was realised early that, although in the sixth grade all the students had followed the traditional curriculum, by the end of the seventh grade the new mathematics was sufficiently different from the traditional approach, that use of the same achievement tests for both groups would be inappropriate. The picture was further complicated by the fact that two similar, but competing, new mathematics curricula had been introduced in the seventh grade, and students of both appeared with sufficient frequency in the sample to warrant a separate analysis. While at Grade 7 these curricula were similar enough to permit the use of single achievement test it was considered imperative at Grade 8 to have separate tests for the two groups. During Grade 9 a major split of the pupils into academic and vocational streams occurs and it was also thought necessary to have separate achievement tests for each stream. A large number of mathematics tests had therefore to be constructed from the item pool. At each grade level all the tests used contained some items in common (items which were not considered to be particularly associated with any one curriculum). Furthermore, each of the tests at the seventh, eighth and ninth grades contained some items that had been used in the preceding year. In this way it was possible to obtain a complete calibration of the items in the bank (several hundred items were involved). This led to a stable and precisely estimated scale of mathematics achievement in wits, and it was possible to estimate for each child where he stood on this scale at the conclusion of each school year.

When the results are summarised one obtains the average achievement of sampled students studying under each regime as presented in Figure 5 . It is clear that neither new mathematics programme is producing results strikingly better than those of the traditional curriculum, and also that 'New Mathematics A' is a comparative failure, especially in Grades 7 and 8. You will note, of course, the considerable differences between the three groups when the Grade 6 testing took place. This results from the fact that the introduction of middle schools into particular regions of the country was a matter of political expediency rather than random choice. The first middle schools to be established were generally clustered in areas of low educational achievement apart from a group in a somewhat privileged area around Jerusalem. (The proponents of 'curriculum A' were given to the Jerusalem schools for trials of their programme). This accounts for the substantial differences between the groups that one finds at the beginning of the study. It is a fact of life that educational innovations are rarely tried out on random samples of the

population, and the strength of this methodological approach to monitoring is that it appears to handle differences in initial ability and subsequent curriculum content without recourse to the rather dubious procedures of covariance analysis.

FIGURE 5 - Average Achievement at the End of the School Year for Groups of Pupils Studying Three Different Curricula

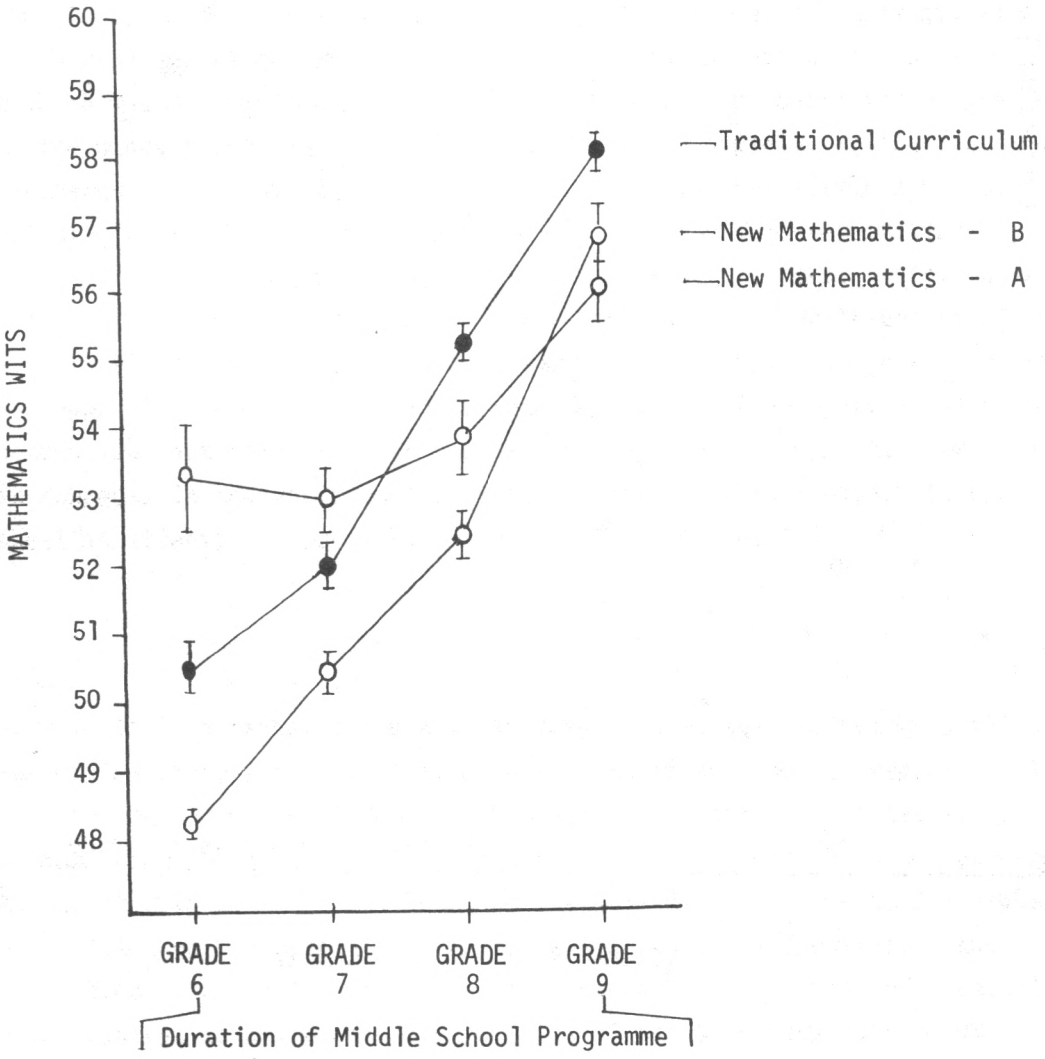
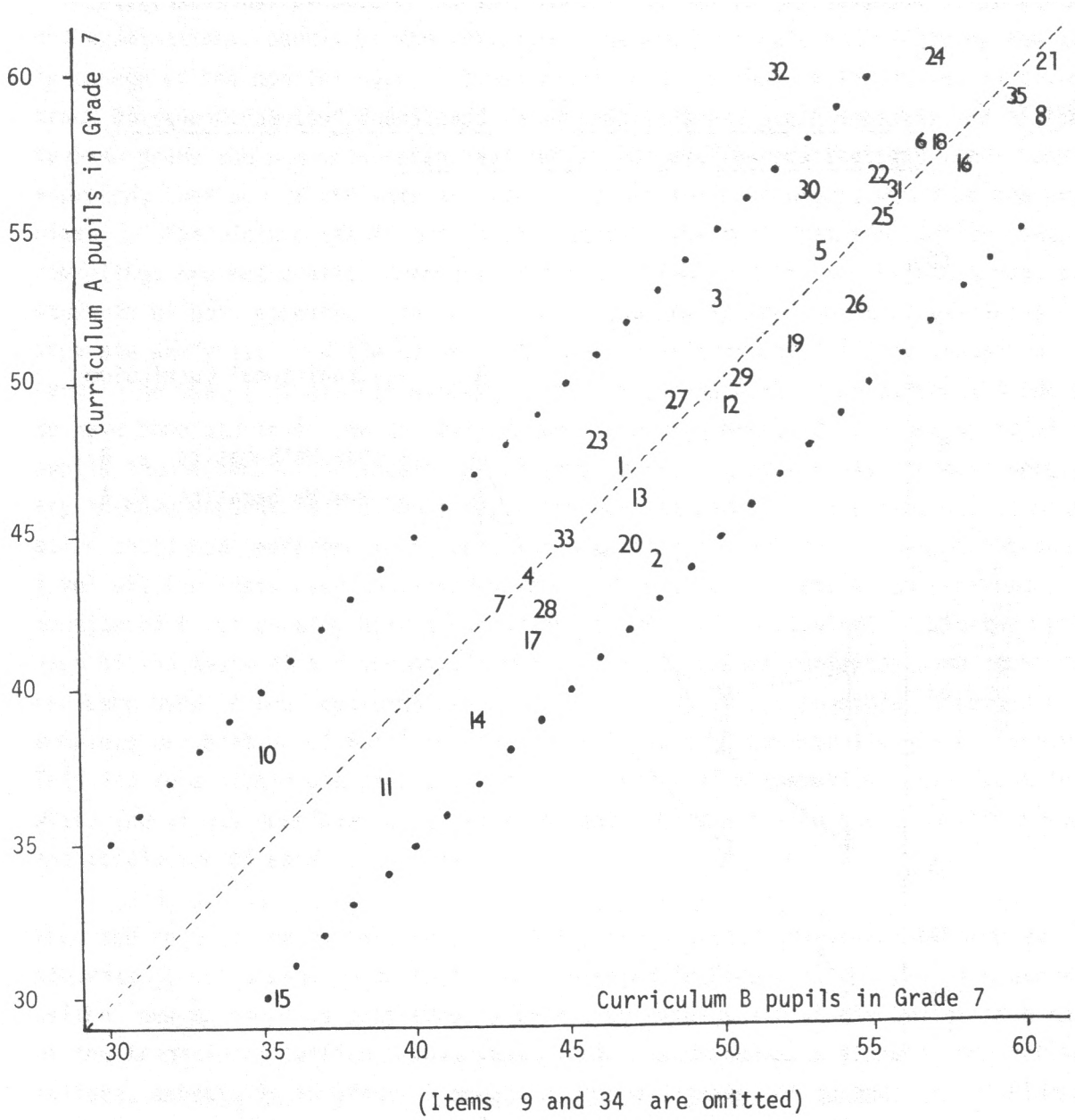


FIGURE 6 – Relative Difficulty of Items for Pupils of Two New Mathematics Curricula



(Items 9 and 34 are omitted)

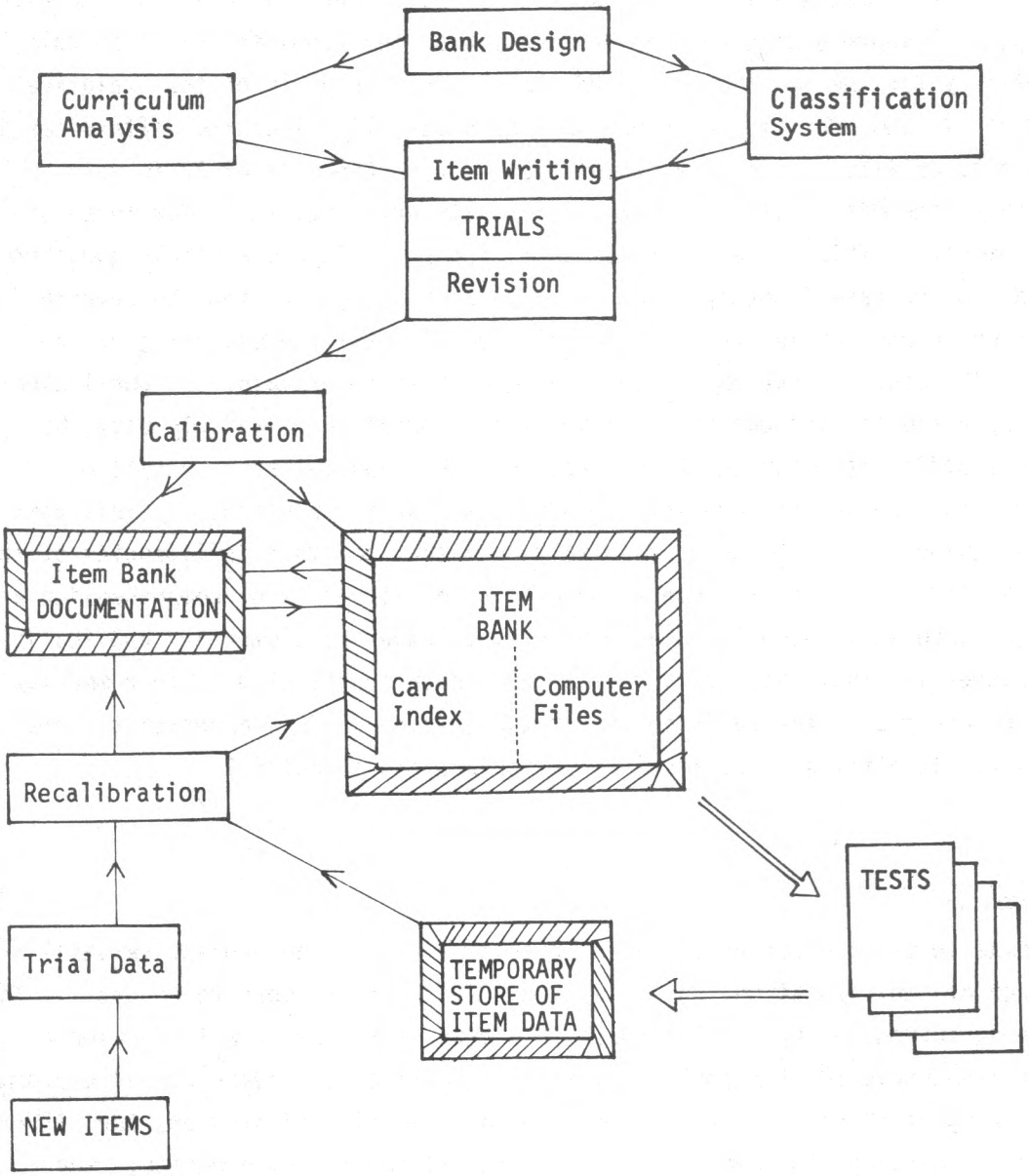
These results, while somewhat discouraging for the proponents of 'curriculum A', are only the beginning. It was important to discover which components of the curriculum lead to such massive discrepancies in learning during Grade 7. At the end of Grade 7 the same mathematics test was given to both 'curriculum A' and 'curriculum B' students. Figure 6 displays the calibration of the items in this test calculated separately for the two types of students. This graph is of the 'relative difficulty' of items for the two groups and automatically discounts differences in the student's ability. It is apparent that, for instance, item 32 is very considerably more difficult for 'curriculum A' students that it is for those of 'curriculum B'. This in fact was not suprising since it was a simple question in 'Cartesian Co-ordinate Geometry' that 'curriculum B' taught during the seventh grade but which was not introduced by'curriculum A' until the beginning of the eighth. The other discrepancies were rather harder to explain. 'Curriculum A' students appeared to find questions involving computation relatively easy, but found great difficulty with questions dealing with graphical representation, algebraic forms and numeric problems requiring algebraic solutions. This work is still continuing. The Grade 8 results reveal that the main deficiencies of the students of 'curriculum A' are concerned with problems requiring comprehension and analysis as opposed to Bloom's categories of knowledge and elementary application. It is hypothesised that this stems from a failure in 'curriculum A' to communicate some basic concepts in the early months of the course. This approach to considering the specific effects of particular curricula appears fruitful.

APPLICATION 3

My last example is in a rather different category in that the project is still at an early stage of implementation, and I have no results of any sort to report. The scheme being introduced by the Indonesian Ministry of Education will develop a system of item banks for the comprehensive assessment of secondary school performance, the monitoring of changes in curriculum, and the selection of students for higher education. Although implementation is now restricted to Mathematics, the Sciences and 'Bahasa Indonesia' (the national language), it is intended eventually to extend the concept to all examinable subjects of the secondary curriculum.

Figure 7 gives an outline of the activity within the project. Three distinct cycles can be identified. The first concerns the design and building of the bank (involving item writing, trials, etc.), and leading to a series of calibration exercises. The next and simplest cycle concerns the generation of tests with specific characteristics, and the feedback to the bank of test results to be used

FIGURE 7 — Projected Item Banking System for Indonesian Ministry of Education



in later re-calibrations. The third cycle concerns the creation of new items and their incorporation into the item bank.

Indonesia is a very large country, comparable in size and population with all of Western Europe, and in consequence there have been serious difficulties in operating a system of national examinations which it is hoped that the projected item bank will overcome.

The first of these difficulties is a language problem. Although there is a national language, there are also about thirty 'local' languages of importance in different regions of the country. In many of the nation's schools 'Bahasa Indonesia' is not yet used as the normal language of instruction, yet hitherto officials in Jakarta have seen no alternative to the conducting of 'national' examinations in the 'national' language. The flexibility of the item banking system will permit different students to be given different examination questions. This opens the possibility of generating examinations in local languages as well as in the national language, so that students whose training in the national language is limited may be more fairly assessed.

Secondly, the difficulties in communicating with remote island provinces, and the extreme pressure for places within the small university sector, have led to certain abuses of the examination system at the local level. The most frequent source of trouble has been the copying and distribution of examination papers prior to the date of the examination itself. With an item banking system which can generate many versions of an examination (such that the pupils in a given classroom may all be responding to different questions) it is hoped that this cheating will be easier to control.

This is an ambitious project on a very large scale and which has still a number of difficulties to overcome before it will have proved its worth. It is operating at the frontier of our psychometric experience and yet I hope that if successful it may be a model for future assessment schemes elsewhere in the world.

REFERENCES

COMBER, L.C. and KEEVES, J.P. (1973). Science Education in Nineteen Countries: An Empirical Study. New York: Wiley.

GULLIKSEN, H. (1950). Theory of Mental Tests. New York: Wiley.

SCRIVEN, M. (1967). 'The Methodology of Evaluation' in Tyler R.W. et al Perspectives in Curriculum Evaluation. AERA Monograph 1: Chicago, Rand McNally (page 58).

WOOD, R. and SKURNIK, L.S. (1969). Item Banking. Slough: N.F.E.R.