CHAPTER 3

Instrument Construction, Data Collection and Processing

The aim of this chapter is to describe briefly the construction of the instruments. A very full description is given by Husén *et al.* (1967, *Volume* I) of the construction of the mathematics tests, questionnaires and occupational classification scheme, and the reader interested in further details is advised to refer to that publication.

Mathematics Tests

In order to formulate the general plan of the tests and the detailed specifications in terms of which they could be constructed, the following steps were taken, as described by Thorndike in Husén *et al.* 1967:

1. The research centre for each participating country was asked to recruit a committee of mathematics educators who would prepare a statement describing the content and objectives of mathematics education in that country.
2. These statements, so far as they were in fact prepared, were examined by a working committee of mathematicians and mathematics educators from several participating countries, and a topical outline was prepared covering the topics that appeared in the reports from the individual countries.
3. The outline was circulated to all participating countries, requesting judgements of the extent to which each topic was indeed covered in the mathematics instruction of the country.
4. On the basis of the responses, together with the judgement of the working committees, simple integral weights were assigned to indicate the importance and emphasis to be given to each topic.
5. In addition to preparing an outline of topics to be covered, attention was given to the types of intellectual processes to be covered.

6. The working committee developed plans relating to the number, length and types of test exercises to be included.

Each National Research Centre organized one or more committees to carry out a content analysis of what was taught in the various grades between Population 1b and the pre-university year, and in some cases the analysis was carried out by school type within a country. The work consisted mostly of an analysis of text books, examinations and teachers' statements. The documents produced by each National Centre were then sent to the International Mathematics Committee.

Two initial outlines were constructed, one for Level 1 (i.e. Populations 1a and 1b) and one for Level 3 (i.e. Populations 3a and 3b). Each outline contained about 40 different topics. A list of the topics for each level is given in Tables A2 and A3 in the Appendix. In each case, however, the objectives or categories of intellectual process were the same, namely:

A. Knowledge and information: definitions, notation, concepts;
B. Techniques and skills: solutions;
C. Translation of data into symbols or schema and vice versa;
D. Comprehension: capacity to analyse problems, to follow reasoning;
E. Inventiveness: reasoning creatively in mathematics.

In Tables A2 and A3 in the Appendix, the column headed *Objectives* indicates the categories of intellectual process that the working committee thought might be appropriately tested in connection with the various topics. The *Importance* column indicates the relative weight to be given in the final testing to each of the topics. (3 signifies great weight, 2 intermediate weight and 1 the least weight.)

Before preparing a pool of test exercises, the mathematics committee had to decide on the length, structure and format of the tests. Three to four hours of testing was accepted as a practical compromise between a comprehensive coverage and what represents a tolerable burden on the time of students and teachers. It was agreed, somewhat reluctantly, to keep the single problems brief. Much as one might like to explore the students' ability to work through an involved sequence of steps, or to develop a complex proof, this seemed not to be possible. Such a task would exhaust too large (and too variable) a fraction of the limited time that was available. Thus, it was decided to limit the tasks to those that a student could be expected to deal with, if he could handle them at all, in not more than, and usually a good deal less than, five or ten minutes for each item.

The requirement of objectivity of scoring suggested the need to fall back on an all-or-none evaluation of a final product—the answer—and this was agreed, not without misgivings, since it was clearly recognised that the restriction placed real limitations on what could be appraised with the test. However, the decision seemed inevitable for an international study involving over a hundred thousand examinees. Furthermore, it was agreed to use mostly multiple-choice type items where the answer choices are supplied and the examinee chooses the best or correct answer. The committee recognised that there are many situations in which producing the response, rather than recognising it, is an essential part of the ability being tested. However, the practical necessity of speeding the scoring of the many papers called for machine scoring and for as extensive a use of multiple choice questions as seemed reasonable within the limits of effective measurement. In the end, 30 of the 174 items in the series required the examinee to write in his answer to a problem while 144 items were in multiple choice format. Using multiple choice items also had the advantage of allowing students to fill their answers in directly on to an IBM 1230 answer sheet which, with very little extra coding at the research centre, could be scored mechanically.

National research centres and members of the test committee supplied illustrative items for each of the topics in the test specifications. Using these items, and also items made available by the Educational Testing Service and by the University of Chicago Examiner's Office, a pool of some 640 items was assembled. Items were selected from this pool and 24 trial test forms were produced; the more elementary forms contained about 22 to 25 items and the more advanced forms 10 to 16 items. Each form was of such a length that it could be easily completed within 45 to 60 minutes. Two anchor items were included in all tests.

The trial test forms were then circulated to National Research Centres, and it was at this point that, as a result of criticism from England, additional trial forms were prepared. Finally, there were twenty-eight trial forms consisting of 497 items. The objective in

preparing the trial forms was to make them inclusive, so that information might be obtained on a wide range of topics and formats.

Each trial form was then translated into the various languages, checked, and pre-tested on judgement samples of about 100 to 150 students in each country. Each test was pre-tested in at least three countries; the assignments were rotated so that different combinations of countries took each of the tryout forms. In each country eight or ten forms were pre-tested. According to the level of the test, it was tried out at the 13-year-old or pre-university level. In some, but not all, countries, appropriate tests were tried out at the 15/16-year-old level.

An item analysis was then carried out in the National Research Centres. Basically, this consisted of calculating the difficulty and discrimination indices estimated by Flanagan's procedure, for each item for a particular sample and reporting these back to the Test Editors. The results from all countries were then entered on to master tables.

The international test committee (Test Editors and Mathematics Educators) agreed that it was desirable to have some parts of the test common to the testing at the four different levels:

(a) 13-year-olds, and the grade group containing the largest fraction of 13-year-olds
(b) an intermediate age or grade group of roughly 15 or 16
(c) a group in the final year of secondary education, but not in a programme with mathematics as a major subject of study
(d) a group in the final year of secondary education *with* mathematics as a major subject of study

It was decided to organize the test in nine one hour units, each of which would be printed in a separate booklet and each of which would constitute a separate "test". The tests taken by each of the populations have already been given in Chapter 2 (see page 29). The items, 174 in all, were selected on the basis of their content validity to the test specifications and on their statistical attributes. In planning the content of the final tests, the editors attempted to maintain a balance between conventional content of mathematics and the newer topics that are being introduced in at least some of the participating countries.

Table 3.1 groups the items into topics in any one set of tests. In

Table 3.1. *Summary of content of tests for different populations.*

Topic	Popn. 1	Popn. 2	Popn. 3a	Popn. 3b
Basic arithmetic	13	3		3
Advanced arithmetic	18	7	3	9
Elementary algebra	12	6	1	5
Intermediate algebra	4	16	19	13
Euclidean geometry	13	17	5	13
Analytic geometry	1	4	8	5
Sets	4	3	4	4
Trigonometric and circular functions		1	3	3
Analysis			8	1
Calculus			9	
Probability			1	1
Logic		2	8	1
Affine geometry	3			

the final analysis, however, seventeen different sub-scores were calculated.

Estimates of the reliability of the total test and subscores were obtained for each population in each country, using the Kuder Richardson procedure of estimating reliability from item statistics and the standard deviation. Formula 20 was used.

Table 3.2 on page 44 gives the reliabilities for the Total Mathematics Score in each country for Populations 1a, 1b, 3a and 3b.

Although the analyses in this book are mostly concerned with Total Mathematics Score, it is of interest to comment on the various groupings of items. Firstly, they were classified, by the pooled judgement of several judges, into items calling for higher mental processes and those calling for lower mental processes. Lower mental process items are those which call for relatively routine application of previously learned techniques, while higher mental process items call for a greater amount of ingenuity and inventiveness in the attack upon novel or complex problems. A second subdivision of the items was into those that consisted of verbally formulated items, in contrast with those that involved primarily computation and solution of a problem expressed in numbers or symbols. A third sub-grouping of items consisted of those which were judged by the mathematics educators to represent the "new mathematics". Fourthly, items were grouped by content areas, i.e., arithmetic, algebra, geometry, etc.

Table 3.2. *Reliabilities of the total mathematics score for populations 1a, 1b, 3a and 3b in each country.*

Country	1a	1b	3a	3b
Australia	.913	.882	.867	—
Belgium	.929	.913	.906	.836
England	.951	.958	.923	.895
Fed. Rep. of Germany	—	.897	.848	.800
Finland	.888	.901	.865	.844
France	.929	.927	.913	—
Israel	—	.917	.817	—
Japan	.941	.941	.925	.926
Netherlands	.948	.915	.794	—
Scotland	.933	.940	.861	.844
Sweden	.869	.869	.897	.732
U.S.A.	.909	.906	.915	.844

Some statistical evidence was gathered on the validity of the IEA tests in England by comparing "O" and "A" Level students' performance on the IEA tests with their performance two or three months later in their "O" and "A" Level examinations. The average correlation was 0.65 for "O" Level and slightly higher for "A" Level, which indicates that there is substantial overlapping, but that it is far from complete. However, in the absence of information on the reliability of the G.C.E., it is not possible to state how nearly the IEA tests and the G.C.E. are measuring the same achievements.

Questionnaires

It was decided to collect information about as many relevant variables as possible that were likely to affect the mathematics performance of the students in the various countries. Among the most obvious factors are home, school and the structure of the educational system. The information about these environmental fields was collected from four main sources: the student, the mathematics teacher, the school principal and an expert on the educational system of each country. Accordingly, there were four types of questionnaires: a Student Questionnaire (ST 1 and 2), a Teacher Questionnaire (TCH 1), a School Questionnaire (SCH 1), and a National Case Study Questionnaire.

The data for variables on the students' background and schooling, collected by means of the Student Questionnaire, concerned such information as grade, sex, age, size of mathematics class, amount of mathematics instruction and homework, father's and mother's occupation[1] and education, aspirations and expectations for further mathematics, further schooling and occupation, best and least liked subjects, examinations taken and extra-curricular mathematics activities. The information requested from teachers concerned mainly teacher certification both in subject matter and professional training, teaching experience, recent in-service training, experience in "new mathematics" and teacher freedom. The information on school characteristics collected concerned school enrolment, number of male and female full-time teachers, number of trained mathematics teachers, type of school, the amount of educational expenditure, age range of students in school and school finance. The National Case Study Questionnaire[1] attempted to collect both quantitative and qualitative data concerning the students in full-time schooling according to school type, selection processes, compulsory schooling, economic data to determine the degree of economic, industrial and technological development and sociological data to determine the role of women in society. This latter questionnaire was completed by one person in each country who not only knew his own system well, but also had a good knowledge of other systems of education.

Only the Student Questionnaires were pre-tested. They were administered (at the same time as the mathematics trial forms were administered) to judgement samples of between 100 and 150 students in each country at both the 13-year-old level (ST 1) and the pre-university level (ST 2). Few modifications proved necessary. The Teacher, School and Case Study Questionnaires were not pre-tested but subjected to comments from experts in the field of questionnaire construction. Research Centres could, if they wished, add extra questions to the questionnaires for the purposes of a national survey.

It was, in some cases, necessary to adapt and modify certain ques-

[1] The construction of an occupational scheme is discussed in detail in Husén *et al.* (1966, Volume I, Chapter 8). Paternal occupation was chosen as the main indicator of family status. Nine categories of occupation were arrived at and agricultural occupations were given special categories within the nine. The difficulties involved in arriving at a classification scheme which is also a scale in all countries were formidable, but it was achieved in a limited way.

tions to national conditions so that a question was comprehensible to those answering it, or so that the information collected was comparable and thus more accurate than a mere translation of the international question; similarly, the source of information varied from country to country for some questions. Thus, for example, in some countries, the head teacher was able to give the data on teachers' salaries, but in other countries this information had to be collected from central records. Examples of the different ways in which the question concerning the extent to which ability grouping was practised within schools are given in Chapter 6.

The coding and punching schemes for the international questionnaires were drawn up by an international committee and these appear in Husén *et al.* (1967) as an Appendix to Volume I. The establishment of international codes was an extremely difficult exercise; the establishment, for example, of one common code into which all school types from all countries could be fitted proved much more difficult than expected, and much discussion and correspondence was required before all were satisfied with and understood the international codes. It should also be pointed out that a Student Opinionnaire was constructed, consisting of two environmental description instruments and five attitude scales, but since none of the data from the Opinionnaire are used in this book, its construction has not been described here.

Data Collection

Administration

It was extremely important to ensure that as far as possible uniform methods of procedure were employed in the testing programme in all countries, and also that very strict standardised procedures were used at the coding and punching stage. In order that this should be the case, a small committee prepared three manuals for National Centres' use. Manual 1 was designed to provide an adequate guide to National Centres concerning all the main procedures to be taken. It included a list of decisions to be made by National Centres, as well as suggestions for sub-sampling within schools and translating and printing the instruments; explanations of particular questions and their codes were also given, as well as instructions for sending all materials to the computing centre. The object was to indicate various methods of procedure to the National Centres in the field work,

and a uniform method of procedure at the coding and punching stage.

Manual 2 was a manual designed for the person responsible for the overall testing programme *within* any one school. The National Centre could decide whether or not it wished to use this in its original or modified form. This manual included a general account of what the project was, the timetable for testing (which varied from country to country), instructions concerning the receiving and storage of testing materials and preparation for the testing sessions, instructions concerning the lay-out of the testing room and the number of invigilators (proctors) required and the briefing of the test administrators and instructions concerning the return of all materials to the National Centre.

Manual 3 (which, again, could be used by the National Centre if so desired) was for test administrators and was the normal type of manual of instructions for test administration. If a National Centre desired to use Manuals 2 and 3 in a modified form, their proposed changes had first of all to be confirmed with the Technical Director.

The total testing programme comprised one and a half days' testing; this imposed a burden on a school, and for those schools where students at different levels were being tested, this burden was considerable. In some of the countries no national survey of this kind had previously been undertaken. This was, therefore, a first experience in large-scale test administration for some National Centres and for the schools, teachers and students in those countries. Difficulties were, of course, experienced, but the results of the experience were encouraging in that few data were lost because of difficulties met in the administration process. It was interesting to note that some National Centres, in whose countries answer sheets had not previously been used, decided to use them. The operation turned out successfully and no difficulties were experienced; the instructions given in Manual 3 on how to fill in the answer sheets appeared to be clear and comprehensive. Apart from the manuals, further instructions were sent out in circular letters, and the main points were every so often summarised in bulletins.

In most cases, the testing in the classrooms was carried out by teachers, but there were exceptions; for example, in Belgium members of the psycho-socio-medicaux centres who are trained in test administration were employed. In Finland, members of the Depart-

ment of Educational Research of the University of Jyväskylä each took responsibility for the schools in a particular area. The department supplied them with cars, and they completed the testing programme within two weeks.

Data Recording

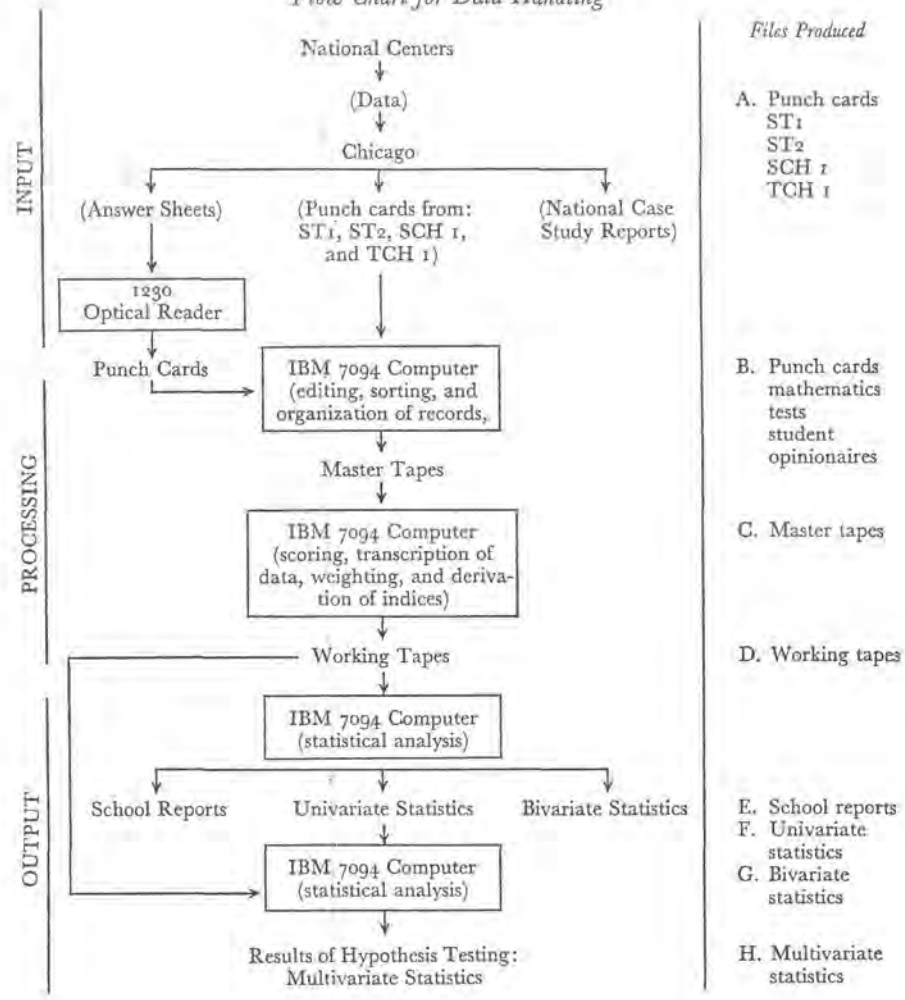
The material from each school was sent to the National Centre. For the laborious and painstaking work of recording the data from the questionnaires on to punch cards or on to special answer sheets designed for the IBM 1230 machine (which then produced a punched card for each answer sheet), each National Centre either employed some of its own staff or hired special staff to do the coding, all of whom worked under the supervision of the person responsible for the IEA project in each centre. Certain questions were asked in different ways in different countries, and it was, as has already been pointed out, of paramount importance that the information given in response to each question was recorded in a standardised way from country to country. For this reason, the responses to as many questions as possible were pre-coded. Where postcoding was required, the columns and ranges on columns (i.e., number of punch positions) were specified.

To ensure standardised recording of data, certain check procedures were set up, which involved National Centres sending their own coding and punching scheme for checking to the IEA secretariat. After this had been approved and when coding and punching had begun at the National Centre, the first twenty punch cards of each type of questionnaire, plus copies of the questionnaires, were sent to Chicago for checking.

National Centres were informed of any errors picked up in these two checks and were asked to correct them before coding and punching of all the questionnaires proceeded.

After all the answer sheets and punch cards were ready, they were despatched to Chicago where all data were entered on to magnetic tape at the University of Chicago Computation Center. When it is realised that in all twelve countries together, 132,775 students from 5348 schools were tested, and that questionnaires were filled in by 13,364 teachers and 5348 headteachers, it will be appreciated that the amount of time required to record these data at National Centres was enormous.

Flow Chart for Data Handling



Data Processing

Although the first data arrived in Chicago in September, 1964, programming had already been underway for a good nine months. The main programmes to be written (apart from programmes for specific hypothesis testing) were the editing, sorting and filing programme, and the programme for compiling the working tapes from the master tape. On the arrival of the Answer Sheets, there was a considerable delay, since it turned out that about one-fifth of all the Answer

Sheets had to have their responses "re-blackened", and a certain number of Answer Sheets had to be completely recopied, since their edges had been damaged in transit.

The data (approximately fifty million pieces) were entered on to the master tape in their raw form (i.e. every response to every item by every individual—student, teacher, head teacher and national case study expert—at every level in every country). Four edited working tapes were compiled, one for each population. All mathematics scores were weighted (see Chapter 2) and corrected for guessing on the working tape, and mathematics sub-scores and various derived indices have been produced. Analyses were then carried out in two stages: first, univariate and bivariate statistics were produced for each population in each country; second, specific hypotheses were tested, as well as a multiple regression analysis being run. The computer used throughout was an IBM 7094. The flow-chart on page 49 may be useful in understanding the total processing system.

Summary

The steps taken in the construction of the mathematics tests were:

(a) content analysis of mathematics courses and statement of objectives of mathematics
(b) preliminary outline of topics and objectives drawn up as test blue-print
(c) topics weighted and test blue-print produced
(d) four hundred and ninety-seven trial items formed into 28 pre-test forms
(e) fourteen pre-tests tried out at 13-year-old level and fourteen at pre-university level on judgement samples of approximately 150–200 students at each level. Each test was tried out in at least four countries. In some countries some tests were also administered to 15/16 year-olds.
(f) item analysis
(g) ten final tests (174 items) constructed such that one test was common to at least two different populations. A maximum of 17 different sub-scores could be computed.
(h) evidence of the concurrent validity of the IEA tests in England was collected for two populations. The average correlation was about .65.

Background information was collected on students by means of a student questionnaire, one version being administered to 13-year-olds (ST 1) and another to the pre-university students (ST 2). These were pre-tested on judgement samples of approximately 100 students in seven countries. Very few changes were required. Background information on the students' teachers and schools was collected by means of a teacher questionnaire (TCH 1) and a school questionnaire (SCH 1). Neither of these was formally pre-tested, but each was worked out by experienced questionnaire constructors. All questions and codes were found to work satisfactorily. Some difficulty was experienced in the establishment of international codes, but it was found that the "common moulds" eventually proved appropriate. Data to provide a contextual background for the findings of the research in terms of the school system and societal and economic factors etc. were collected by means of a National Case Study Questionnaire completed by a national comparative educationist.

Three different manuals were produced for use by National Centres, school testing organisers and actual testers, so as to ensure standardisation of procedure throughout all the full testing programme and coding and punching stages. In most cases, the actual testing was carried out by teachers, but in some cases was carried out by trained testers or by students of psychology or education.

All responses to the mathematics items were recorded on specially prepared IBM 1230 answer sheets. Responses to questionnaire items mostly pre-coded, but some required post-coding) were punched on punch cards at the National Centre, but only after a series of checks had been carried out on the punching of the first twenty of each type of questionnaires. Answer Sheets and punch cards were then sent to the University of Chicago Computation Center and there all responses were entered on to a master tape. Working tapes were compiled, involving the weighting of scores and the derivation of sub-scores and special indices. Analyses were then carried out in two stages—the production of univariate and bivariate statistics and the testing of specific hypotheses.