DOCUMENT RESUME

ED 357 955 SE 053 128

AUTHOR Keeves, J. P.

TITLE Technical Issues in the First and Second IEA Science Studies.

PUB DATE 92

NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 24, 1992).

PUB TYPE Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Comparative Education; *Cross Cultural Studies; Data Analysis; Data Collection; Elementary Secondary Education; Foreign Countries; International Education; International Programs; Models; *Research Methodology; *Research Problems; *Science Education; Testing Programs; Test Results

IDENTIFIERS Educational Issues; First International Science Study; *International Assn Evaluation Educ Achievement; International Evaluation Education Achievement; International Studies on Educational Achievement; *Science Achievement; *Second International Science Study

ABSTRACT

The first and second International Association for the Evaluation of Educational Achievement Science Studies examined and compared science achievement in 19 and 26 countries, respectively. This paper considered 10 technical issues that arose in these 2 studies of science education and achievement. The issues discussed were the: (1) need for the development of a theoretical framework for cross-national studies of educational achievement; (2) need for national indicators of the conditions and outcomes of education; (3) difficulty of determining equivalent target populations in different countries; (4) difficulties related to longitudinal and cross-sectional studies; (5) methods of analysis; (6) units and levels of analysis; (7) problems in scoring the tests and scaling the achievement test data; (8) sample design and execution; (9) problems involving sampling errors and significance testing; and (10) translation and national variations in instruments. Appended information includes a list of 15 references; tables of sizes of achieved samples and response rates for the two studies; and schematic models of the context and components of the science curriculum, performance in science, and science achievement scale. (MDH)

TECHNICAL ISSUES

IN THE

FIRST AND SECOND IEA SCIENCE STUDIES

J.P. Keeves
School of Education
The Flinders University of South Australia
Bedford Park, South Australia
Australia, 5042

Paper presented at AERA Annual Meeting
San Francisco, Friday 24 April 1992
Session 52.15 *Technical Issues in International Assessment*

TECHNICAL ISSUES IN THE FIRST AND SECOND I.E.A. SCIENCE STUDIES

Introduction

The First I.E.A. Science Study was undertaken as part of the I.E.A. Six Subject Study and was constrained by the fact that it was conducted simultaneously with the studies of Reading Comprehension and Literature. These studies were tightly controlled from the International Coordinating Centres in Hamburg and later in Stockholm.

The Second I.E.A. Science Study was planned as a 'do it yourself study' and as a cooperative research program. Only in this way was it possible for such a substantial project that initially included 30 or more countries to proceed with external funding that was limited to a mere US$120,000 to cover the international costs of coordinating the study. The six meetings held once a year from 1981 to 1986 of the International Study Committee were central for the training of the National Research Coordinators and the planning and conduct of the cooperative program of research. The emphasis in the first study was on the preparation of the international reports, while in the second study the emphasis was on the preparation of national reports and subsequently on the international analyses and the international reports.

The first study involved 19 countries. In the second study 26 countries tested students in science. In Mexico and Tanzania the research institutes conducting the study were closed in 1984, and their data were lost. In Canada (French speaking), Ghana and Zimbabwe, the research centres were either closed, or key staff had their appointments terminated, so that only by the valiant efforts of individuals and support from agencies providing foreign aid was it possible to salvage the data for the preparation of the national and international reports.

The sizes of the two studies may be seen in Table 1. In addition to the administration of paper and pencil achievement tests, a practical skills testing program was carried out in each study. In the first study at the 14 year-old level in two countries, and at the terminal secondary level in one country, students were tested. However, in the second study, six countries took part in the practical skills testing, and at the 10 year-old and 14 year-old levels 7,700 and 9,000 students respectively were tested. In general terms the second study was approximately twice as large as the first.

Table 1 Participation in the First and Second IEA Science Studies

First IEA Science Study: 1970-71

	Countries	Schools	Teachers	Students
10 year-olds	16	1917	9310	38,672
14 year-olds	18	2055	8216	48,414
Terminal secondary	18	1580	8241	49,734
Total	19	5552	25767	136,820

Second IEA Science Study: 1983-84

	Countries	Schools	Teachers	Students
10 year-olds	18	3096	5065	81,855
14 year-olds	24	3658	9830	94,974
Terminal secondary	18	2828	7860	85,449
Total	24	9582	22,755	262,276

Table 2 Published Items from the First and Second IEA Science Studies

	First Study	Second Study
International Reports		
Major	1	6
Minor (Cross-national)	2	3
National Reports	7	50
Issues of Journals	-	3
Articles in Journals and Books	25	58
Theses	5	17
Technical Documents	-	2
Symposium Reports	1	1
Secondary Analysis Reports	10	1
Total	51	141

Studies of this magnitude, when carried out with very limited resources, but with a great deal of good will and strong collegiate relationships among the research workers participating in the conduct of the study, encounter many problems. While for many working in a foreign language is a hurdle to be overcome, it seems unrelated to the quality of data collected. However it cannot be denied that there are shortcomings and deficiencies in the data submitted for the international analyses from some countries in both studies. Some of these problems can be surmounted by care in the analysis of the data and some cannot. Nevertheless, these two data sets form the richest and largest body of data on the teaching and learning of any school subject ever collected. The links between the two data sets are strong although not perfect. Either together or separately they are a massive resource for the continuing study of science teaching across the world.

The extent to which the data have already been studied can be seen in Table 2 where the number of separate publications and reports from these two studies of science education, known to have been issued, are recorded. It is easy to draw attention to deficiencies and debate the technical issues involved in the design, data collection, analysis and reporting of these two studies. Likewise, it is easy to be critical after the event and to denigrate the devoted efforts of those involved. However, the magnitude of the output from these two studies must indicate the contribution of these projects to science education and educational research across the world. Moreover, the policy of encouraging countries to prepare their own national reports would seem to have had very beneficial results in terms of the number of items produced and published. From the first study 34 of the 51 items came from only two countries - Australia and Sweden. In the second study 37 and 12 of the listed 141 items were generated in the United States and Canada respectively. It is not surprising that Australia and Sweden should have shown a commitment to the second study, and Canada and the United States to the third study, which is currently being planned. It should also be noted that the further production of secondary analyses, such as those prepared in Stockholm by the Spencer Fellows, is still being carried out for both the second study and for the examination of change over time.

ISSUES

It is of value in retrospect to consider some of the technical issues that arose in these two studies of science education and achievement.

1. A Theoretical Framework for Cross-National Studies.

It might be argued that the development of a theoretical framework for cross-national studies of educational achievement is not a technical problem but a conceptual issue. However, without a theoretical framework the design of the study, the instrumentation, and the analysis of data lack direction and purpose, and some technical problems become difficult, if not impossible, to resolve. The IEA science studies have been wrongly attacked for a failure to develop theoretical perspectives. Two models were advanced for the second science study. The first is the Model of the Science Curriculum which was initially advanced in 1972 during the analysis of the data from the first study (Keeves, 1974). It is shown in Figure 1. The second is a model of performance in science initially presented at the first meeting of the International Study Committee for the second study in 1981 and derived largely from the analyses of data collected in the first study (Keeves, 1984). This model is shown in Figure 2. Nevertheless, there is little theory to guide the cross-national analyses of data. With the larger number of countries and school systems engaged in IEA studies there is an urgent need for comparative educators to develop theory to guide the analysis of data at the system or country level.

2. National Indicators of Conditions and Outcomes

A sustained search was made to try to obtain information on a range of national indicators of the conditions and outcomes of education. Information on indicators considered to be useful for cross-national analyses, for use as marker variables to test the quality of the samples, or as indicators of change occurring over the period under survey were simply not available from many countries. It was, for example, a long struggle to obtain accurate values for the common index of retention rate at the terminal year of secondary schooling, even from a country such as The Netherlands that is relatively advanced in social science research. The only marker variable that could be employed was the male-female ratio, and this was frequently not available for the terminal year of secondary schooling, and very rarely available for the study of particular fields of science. Indeed, the availability of such information would appear to have deteriorated even in the OECD countries between 1970-71 and 1983-84, although perhaps the figures published by Unesco in the early 1970s were grossly in error. Over a period when there was a marked change in the roles of women in society, it would be thought that information on relevant indicators, in addition to the proportion of women engaged in the labour force, would be available.

3. The Definition of Target Populations

The basic target populations for both the first and second science projects were students (a) aged 10 years, (b) aged 14 years, and (c) in the terminal year of secondary schooling. However, in different countries students start at school at different ages ranging from 5 to 7 years, and hence by the age of 10 years have had different periods at school. Moreover, the total number of years of schooling provided ranges from 10 years in the Philippines, to 13 years in England, Canada (Ontario), Ghana, Hong Kong, Italy and Singapore. In addition, countries differ in the number of years for which schooling is mandatory, and after the completion of compulsory schooling in some countries students are allocated to either academic or vocational schools, while in others all students who remain at school continue with an academic or general program.

As a consequence of these differences it is somewhat unsatisfactory to define the target populations in terms of age or number of years of schooling completed or grade level currently enrolled. Further difficulties arise in some countries where there is interest in examining classroom and teacher effects, which together with the administrative convenience to schools of testing complete class groups, lead to a preference for sampling only intact classes. As a consequence, in testing age samples at the 10 year-old and 14 year-old levels, instead of sampling across grades, students must be drawn from only one or two classes at specified grade levels. Clearly strict comparability is not possible, because both age and grade have effects on student achievement.

Likewise, at the terminal secondary school level,by specifying the requirements of attendance in full time normal schooling, and studying courses that lead to entry to higher education, difficulties associated with attendance at vocational schools are removed. However, the re-enrolment of adult students in upper secondary courses must increasingly be considered. Further difficulties arise from an interest in testing students specializing in the study of one or more of the major fields of science, biology, chemistry and physics, or studying a more general type of science, or a non-basic science related field such as astronomy, oceanography or environmental science. If care is taken in defining the target populations to answer the research questions of importance within a country and so that general comparability is also achieved across countries, then this is probably the best solution to a difficult problem.

A further complication arises when after examination of the tests to be employed the research workers in a particular country realize that the tests, which are appropriate in other countries, are too difficult for the comparable grade in their country. As a consequence they argue that it is necessary to test at a higher grade level. This is useful information and can at least to some extent be taken into consideration in the reporting and discussion of results..

If the investigation were an international olympics or horse-race with a prize to the top nation, then these would be insuperable problems. While there is always interest is comparability of performance across countries, the science educator is more interested in the patterns of results, and the magnitudes of effects after appropriate allowances have been made, rather than the strict statistical significance of the differences in achievement between countries. Moreover, for the research worker there is interest in why differences are observed across countries and in relationships rather than point estimates.

4. A Longitudinal or Cross-Sectional Study

The first science study was a cross-sectional study at three age levels. Consideration was given as to whether a replication study should be carried out or whether the study should examine change in performance over a school year at the classroom and student levels of analysis. This issue had also been raised in 1966 when the first study was being planned, and was examined with data at the time of planning the second science study. No way was known in 1981 of handling the many problems of: (a) examining simultaneously both individual and group effects; (b) bringing the pretest and posttest data to a common scale with sufficient accuracy that change for individual students and for classroom groups could be examined effectively; and (c) measuring effectively in a survey study classroom practices that might influence change in achievement over a school year. It was decided that until measurement and analytical procedures were known to be available that would yield sound findings, it was inappropriate to orient the study towards the examination of change over a school year in the 30 or more countries that would possibly be involved.

The emphasis of the second science study was then oriented towards the investigation of change over the 14 year period from 1970-71 to 1983-84, which as we thought represented seven rich years and seven lean years. Thus a longitudinal study of school systems was planned and both changes in outcomes and conditions of science teaching and learning were investigated. To provide some control for the effects of prior learning both on teaching conditions and attitudes, two measures of aptitude for the learning of science, namely a word knowledge test and a computational skills test, were included in the battery of instruments employed. The problem of the investigation of individual and group change in classroom and teacher studies where performance is measured at only two points in time, to my knowledge, has not as yet been shown to be feasible and highly rewarding even within one country.

5. Methods of Analysis

The major problem in the analysis of educational survey data was argued out in IEA around the conference table in Febuary 1965, and a major change in plans for the analysis of data was made at that time. Even though it was initially planned that multivariate analysis of covariance procedures should be used for the analysis of data in the First IEA Mathematics Study, a shift was made to the use of regression analysis procedures. The basic problem is that home background, prior performance and aptitude not only influence final performance but also attitudes, motivations and the school and classroom conditions provided for learning. Unless allowance is made for this in analysis the results obtained are grossly in error. The different methods of analysis: (a) stepwise regression, (b) path analysis, (c) factorial modelling, (d) canonical correlation analysis, (e) partial least squares path analysis, and (f) LISREL, were systematically examined with the same data set and the results published (Keeves, 1986). The problems associated with the analysis of multilevel data were also thoroughly examined (Keeves and Lewis, 1983; Larkin and Keeves, 1984) and subsequently the range of possible solutions to the question systematically investigated, and an initial approach involving the use of a path model developed (Cheung et al. 1990)

Partial least squares path analysis was chosen as a flexible and readily applied analytical procedure for testing the path model which was advanced in the design of the study. Some analyses were carried out with LISREL, but its use other than in a limited way was rejected on several grounds. First, it cannot provide appropriately for the inclusion of dichotomous variables such as sex of student in a maximum likelihood analysis except by the simultaneous analysis of single sex groups. Secondly, it does not permit endogenous latent variables to be formed by other than the reflective mode, as is appropriate for a factor analytic approach. Thirdly, LISREL is heavily dependent on the statistical significance of the measures estimated in analysis. However, the levels of significance provided by the LISREL program are statistically erroneous where cluster samples have been used. There was no way known at that time within LISREL to model the hierarchical nature of the data in order to obtain more appropriate estimates of error and levels of statistical significance.

Hierarchical Linear Modeling was found to be a very promising technique, but lack of time prevented its use with the many large and complex data sets available for analysis. It has since been used effectively in the secondary analyses of data from the second study (Kotte, 1992). However, it has not proved possible to examine in the one analysis data from single sex and coeducational schools and to investigate the effects of student sex differences in these different educational settings.

6. Units and Levels of Analysis.

In both studies, there was awareness of the problems of aggregation bias when student level data are combined to the group level for analysis, and of the less commonly discussed "disaggregation bias" when group level data are disaggregated to the student level. While HLM provides a promising approach, the amount of data to be analyzed and the complexity of the model being tested prevented its use. A preference was developed for the analysis of data and the testing of path models using partial least squares path analysis at: (a) the between student within group level, and (b) the between group level. However, the level of analysis also depends on the issues being addressed. Many findings reported from the second science study involve some bias, because of an inability to handle effectively the multilevel nature of the data.

7. Scaling the Achievement Test Data

Several problems arose in the scoring of the tests: (a) a scheme of core and rotated tests was employed at all three levels; (b) students were requested not to guess blindly and as a consequence some correction for guessing needed to be employed to allow for substantial differences between countries in the tendency to omit a test item rather than guess a response;

(c) there were occasional items that were incorrectly printed; (d) there were occasional items where an error had arisen in translation; (e) the tests needed to be equated across age levels; (f) the tests needed to be equated over time; and (g) there was one case where not all core test items were included in the tests administered to students. The study by Sontag (1983) had shown that with the United States' science data from the first study the one parameter IRT model produced the most stable results for vertical equating. Using the BICAL and MSCALE programs a science achievement test scale was developed and is shown in Figure 3. There were two problems. First, the MSCALE program was found not to do what it was purported to do in the treatment of missing data. Thus the adjustment made for omitted responses, while better than ignoring the problem, was less than satisfactory. Secondly, the second science study yielded student level standard deviations for each country at the 14 year-old level that were approximately 1.3 times greater than the corresponding cr untry standard deviations obtained from the tests used in the first study. However, this did not occur at the 10 year-old level where the standard deviations were remarkably consistent over time for each country, in spite of substantial gains in achievement. It could be that the change in standard deviation at the 14 year-old level arose from differences in the properties of the tests employed on the two occasions.

The issue of whether or not a total test score should be employed was concerned with whether the biology, chemistry and earth science, and physics subscores could be combined. Munck (1979) in a reanalysis of data from the First Science Study for three countries, with different types of science curricula and curriculum control provided strong evidence to support the use of a total score. Likewise the analyses by Peaker (1969) from the First Mathematics Study, together with some principal components analyses and the extensive IRT analyses carried out on data from both the First and Second Science Studies all supported the use of a total score. However, in the PLS analyses, the subtest scores for biology, chemistry and physics could be more effectively combined in the analysis in preference to using a total score.

8. Sample Design and Execution

G.F. Peaker laid down the sampling procedures to be employed in the early IEA studies. Two sampling designs were thus known to work in cross-national studies. Both designs are two-stage sample designs, sampling first by schools and then by students within schools. In the first design a stratified simple random sample of schools is chosen and then with a constant sampling fraction within strata, either all or a specified proportion of students are selected from within schools. In the second design a stratified probability proportional to size of school sample of schools is selected and a fixed number of students, either a class or 25 students, is sampled from within each school.

Problems arise in several situations. First, in countries where a complete listing of schools does not exist, a three stage sample design must be employed with school districts selected at the first stage. Secondly, where intact classes are preferred for testing, since no school likes to test a weak class of students, it becomes necessary for the selection of the class of students to be done at the National Coordinating Centre, and care must be taken to ensure that the class chosen is tested. Thirdly, some major research institutes have their own sampling experts. who in their ignorance, employ variations that introduce serious bias. The most common error is to draw a simple random sample of schools and then to test an intact class from within a school and not to make allowance for the size of the school. Fourthly, it is inevitable that not all schools tested can participate. Likewise, some students selected are generally unable to take part in the testing program. The accepted procedure is to select a matching replacement sample of schools and students, and to replace a school or student with a matched replacement drawn randomly in advance. Some sampling experts within research institutes reject these replacement procedures, and prefer to report low response rates, and accept the possible bias in the samples. However, the use of replacement samples is likely to introduce less bias than ignoring the problem. Fifthly, some National Research Centres hold insufficient information about the factors that influence educational outcomes to be able to stratify effectively the sampling design in order to reduce error and to weight for differential losses. Finally, in countries where participation in a testing program is optional and not mandatory, insufficient effort is made to persuade schools to take part in the study. The bias introduced is both unknown and cannot be allowed for when undertaking analyses at the country, school and student levels. Sample design and execution would appear to be a major shortcoming of research training in many countries. However, the quality of sample design and execution is unrelated to the size of a country, since both large and small countries seem to encounter problems. Tables 3 and 4 record information on sample quality for 10 countries for the first and second science studies respectively. It is possible to see from these tables where some but not all of the low response rates arose as a result of unsatisfactory sample design and execution.

9. Sampling Errors and Significance Testing

Educational research is driven by the emphasis placed in research training on statistical significance testing rather than the magnitude of effects and pattern of results. However, the reported sampling errors employed to test for significance in a very large proportion of cases in educational research are simply wrong, since they make no allowance for the design effects associated with cluster sample designs. These are the rule rather than the exception in educational research.

The samples employed in IEA studies involve two and three stage sampling, are highly stratified, and require weighting to correct both for different sampling fractions across strata and differential losses. The use of a variance ratio is generally unsatisfactory in the estimation of the design effect and is limited to the estimation of the error of a mean. Three more appropriate procedures are available: (a) Jackknifing, which best involves dropping one primary sampling unit at a time; (b) Taylor's series estimation; and (c) Bootstrapping. The first is the most widely used (see Rust, 1984) but routines are not available in any computer package except OSIRIS. From the use of these procedures the sampling error of an estimate can be obtained and from this DEFF (the design effect), ROH (the ratio of homogeniety) and standard errors of the mean can be calculated. DEFF and ROH differ from sample to sample, for each measure, and for each statistic – means, correlation coefficients, regression coefficients etc. To calculate DEFF and ROH for means by jackknifing is a relatively simple and quick task, as it is for correlation and regression coefficients for simple random samples. However, for complex cluster samples, attempts were made to develop a procedure for the second science study to estimate the standard errors of the correlation and path coefficients, without repeating each analysis several hundred times. However, the task was too difficult and such estimates were not obtained. Instead judgments were made that were based on crude estimation of the size of a parameter which was likely to be significant at the 5 per cent level and these judgment values were employed. Work carried out in the First Science Study (Peaker, 1975) and at the Australian Council for Educational Research (Ross, 1976; Wilson, 1983; Farish, 1984) provided guidance in specifying these judgement values. Only for mean total scores for a country were the standard errors calculated, and although these were the best estimates of error that could be obtained, their limited accuracy did not seem to warrant the use of elaborate significance testing procedures for comparing differences between national means.

10. Translation and National Variations in Instruments

Care was taken in the translation of the tests by using double translation and back translation procedures. The very extensive item analyses using both traditional procedures and IRT procedures led to the detection of some translation problems. They were, however, few and far between, and did not contaminate the results to any recognizable extent.

Some problems arose through national variations to instruments. In Japan, for example, the Teacher Unions did not permit questions to be asked of students on father's and mother's occupations. Such problems showed up in the careful analysis of the data and ways to circumvent these problems had to be found.

Conclusion

The data sets are rich, detailed and sufficiently accurate for highly valuable findings to have emerged from both the First and Second Science Studies. Tribute must be paid to Leif Andersson in Stockholm and to Ditmar Jungnickel and his team of willing and highly competent coworkers at the University of Hamburg. They did a remarkable job on the detailed analyses of the large bodies of data that they analyzed from both the First and Second Science Studies, and with great care and commitment. With studies of this magnitude a very large number of people has been involved. IEA thanks them one and all for their contributions to this major international investigation into science education in the years 1970 and 1984. It is now possible and desirable for detailed secondary analyses to be carried out, in order to further an understanding of the teaching and learning of science in a changing world.

INTERNATIONAL REPORTS

First IEA Science Study

Comber, L.C. and Keeves, J.P. (1973) *Science Achievement in Nineteen Countries.* Almqvist and Wiksell, Stockholm and Wiley, New York.

Second IEA Science Study

IEA (1988) *Science Achievement in Seventeen Countries. A preliminary report.* Pergamon, New York and Oxford.

Rosier, M.J. and Keeves, J.P. (1991) *The IEA Study in Science I. Science Education and Curricula in Twenty-Three Countries.* Pergamon, Oxford.

Postlethwaite, T.N. and Wiley, D.E. (1991) *The IEA Study in Science II. Science Achievement in Twenty-Three Countries.* Pergamon, Oxford.

Keeves, J.P. (ed.) (1992) *The IEA Study in Science III. Changes in Science Education and Achievement: 1970 to 1984.* Pergamon, Oxford.

Doran, R.L. and Tamir, P. (eds.) (1992) An international assessment of science practical skills. *Studies in Educational Evaluation,* 18(1), pp.1-102.

Summary Report

Keeves, J.P. (1992) *Learning Science in a Changing World. Cross-National Studies of Science Achievement: 1970 to 1984.* IEA, The Hague.

References

Cheung, K.C., Keeves, J.P., Sellin, N., Tsoi, S.C. (1990) The analysis of multilevel data in educational research. *International Journal of Educational Research,* 14(3), 215-319.

Farish, S.J. (1984) *Investigating Item Stability: An empirical investigation into the variability of item statistics under conditions of varying sample designs and sample size.* ACER, Hawthorn, Victoria.

Keeves, J.P. (1974) The IEA Science Project. Science achievement in three countries – Australia, the Federal Republic of Germany, and the United States. In *Implementation of Curricula in Science Education.* German Commission for Unesco, Cologne. pp.158-178.

Keeves, J.P. (1984) Conceptual framework for science learning *Pedagógiai Szemle*, 23(12), 1170-76.

Keeves, J.P. (ed.) (1986) Aspiration, motivation and achievement: different methods of analysis and different results. *International Journal of Educational Research*, 10(2), 115-243.

Keeves, J.P. and Lewis, R. (1983) Issues in the analysis of data from natural classroom settings, *Australian Journal of Education*, 27(3), 274-287.

Kotte, D. (1992) Gender Differences in Science Achievement in 10 Countries: 1970-71 to 1983-84. Ph.D. dissertation, University of Hamburg.

Larkin, A.I. and Keeves, J.P. (1984) The Class Size Question: A study at different levels of analysis. ACER, Hawthorn, Victoria.

Munck, I.M.E. (1979) *Model Building in Comparative Education*, Almqvist and Wiksell, Stockholm.

Peaker, G.F. (1969) How part scores should be weighted? *International Review of Education*. 15(2), 229-237.

Peaker, G.F. (1975) *An Empirical Study of Education in Twenty-One Countries – A Technical Report.* Almqvist and Wiksell, Stockholm.

Ross, K.N. (1976) *Searching for Uncertainty. An empirical investigation of sampling errors in educational survey research.* ACER, Hawthorn, Victoria.

Rust, K.F. (1984) Techniques for estimating variances for sampling surveys. Ph.D. dissertation, University of Michigan, Ann Arbor.

Sontag, L.M. (1983) Vertical equating methods: a comparative study of their efficacy. Unpublished doctoral dissertation, Teachers College, Columbia University.

Wilson, M. (1983) *Adventures in Uncertainty. An empirical investigation of the use of Taylor's series approximation for the assessment of sampling errors in educational research.* ACER, Hawthorn, Victoria.

Table 3 Sizes of achieved samples and response rates: 1970-71

Numbers	Australia	England	Finland	Hungary	Italy	Japan	The Netherlands	Sweden	Thailand	United States
POPULATION 1										
Numbers:										
Schools	-	162	97	156	264	250	60	98	27	259
Students	-	3556	1290	4860	4508	2467	1622	1982	1822	5431
Teachers	-	1301	350	846	373	1552	166	665	-	1632
Response Rates (%)										
Schools	-	79	99	99	73	100	66	99	94	68
Students	-	73	97	95	49	100	65	96	82	64
POPULATION 2										
Schools	221	144	77	210	327	196	50	95	29	142
Students	5301	3256	2325	7026	7363	1946	1236	2475	1932	3935
Teachers	1638	1498	496	1520	1152	752	267	1157	-	992
Response Rates (%)										
Schools	99	66	100	100	86	98	52	96	90	57
Students	96	60	98	94	83	98	49	91	81	46
POPULATION 3										
Numbers:										
Schools	194	70	77	39	253	--	38	142	13	114
Students	4194	2274	1807	2855	4877	--	1164	2988	724	2600
Teachers	1600	867	630	451	1538	--	179	2131	-	816
Response Rates (%)										
Schools	99	32	100	100	70	--	39	95	95	43
Students	92	27	82	98	61	--	37	90	66	35
POPULATION 3 SPECIALISTS										
Numbers:										
Biology	1357	427	1103	1666	90	--	611	932	614	394
Chemistry	1651	511	412	178	380	--	619	682	599	414
Physics	1622	587	785	2179	2521	--	528	1112	554	618
Non-Science	1640	1218	414	630	1949	--	338	990	841	1311

Sources of data: Comber and Keeves (1973, p.45), Peaker (1975, pp.36-37)

Table 4   Sizes of achieved samples and response rates: 1983-84

Numbers	Australia	England	Finland	Hungary	Italy[a]	Japan	The Netherlands	Sweden[b]	Thailand	United States Phase A[c]	United States Phase B[d]
POPULATION 1											
Numbers:											
Schools	220	181	106	100	119	221	-	134	-	121	123
Students	4259	3698	1600	2590	5152	7924	-	1270	-	2909	2822
Teachers	819	380	121	97	309	543	-	137	-	120	117
Response Rates (%)											
Schools	78	66	96	100	58	99	-	68	-	48	88
Students	67	61	86	95	84	99	-	e	-	16	77
POPULATION 2											
Numbers:											
Schools	233	146	90	99	224	199	224	137	96	88	119
Students	4917	3069	2546	2497	4622	7610	5025	1181	3780	1958	2520
Teachers	1630	1077	244	354	1425	314	393	190	96	88	113
Response Rates (%)											
Schools	84	59	97	99	75	99	92	59	93	34	85
Students	74	58	90	92	78	95	86	e	92	31	69
POPULATION 3											
Numbers:											
Schools	165	127	86	77	317	193	--	214	98	164	100
Students	5057	3737	3638	2001	6848	6561	--	4033	7124	4774	1729
Teachers	700	790	84	177	535	123	--	-	431	-	114
Response Rates (%)											
Schools	80	49	93	96	69	96	--	76	98	33	-
Students	69	41	90	89	75	91	--	59	89	38	-
POPULATION 3 SPECIALISTS											
Numbers:											
Biology	1631	884	1652	301	147	1212	--	619	1171	-	659
Chemistry	1177	892	971	143	217	1468	--	1172	1168	-	537
Physics	1073	917	810	398	1766	1187	--	1156	1168	2719	485
Non-Science	995	1004	-	1036	2455	2230	--	1281	3685	2055	-

Notes:
a. Italy, Grade 8 sample
b. Sweden, Grade 7 + 8 samples combined (14 year-olds only)
c. United States data from the Phase A Study conducted in 1983-84.
d. United States data from the Phase B Study conducted in 1986.
e. No information available to estimate accurately student response rates.

Sources of data: IEA (1988, 81-83).

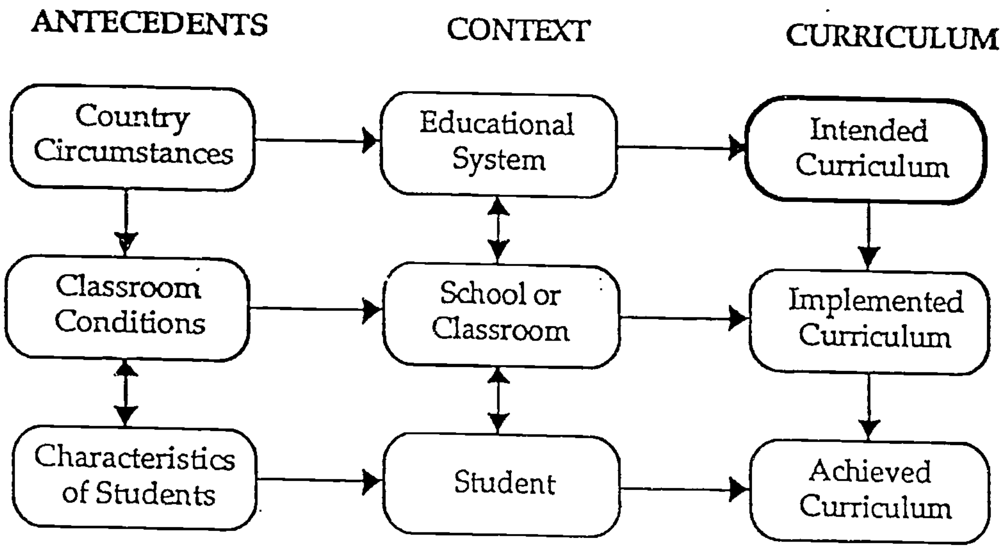

FIG. 1. The context and components of the science curriculum

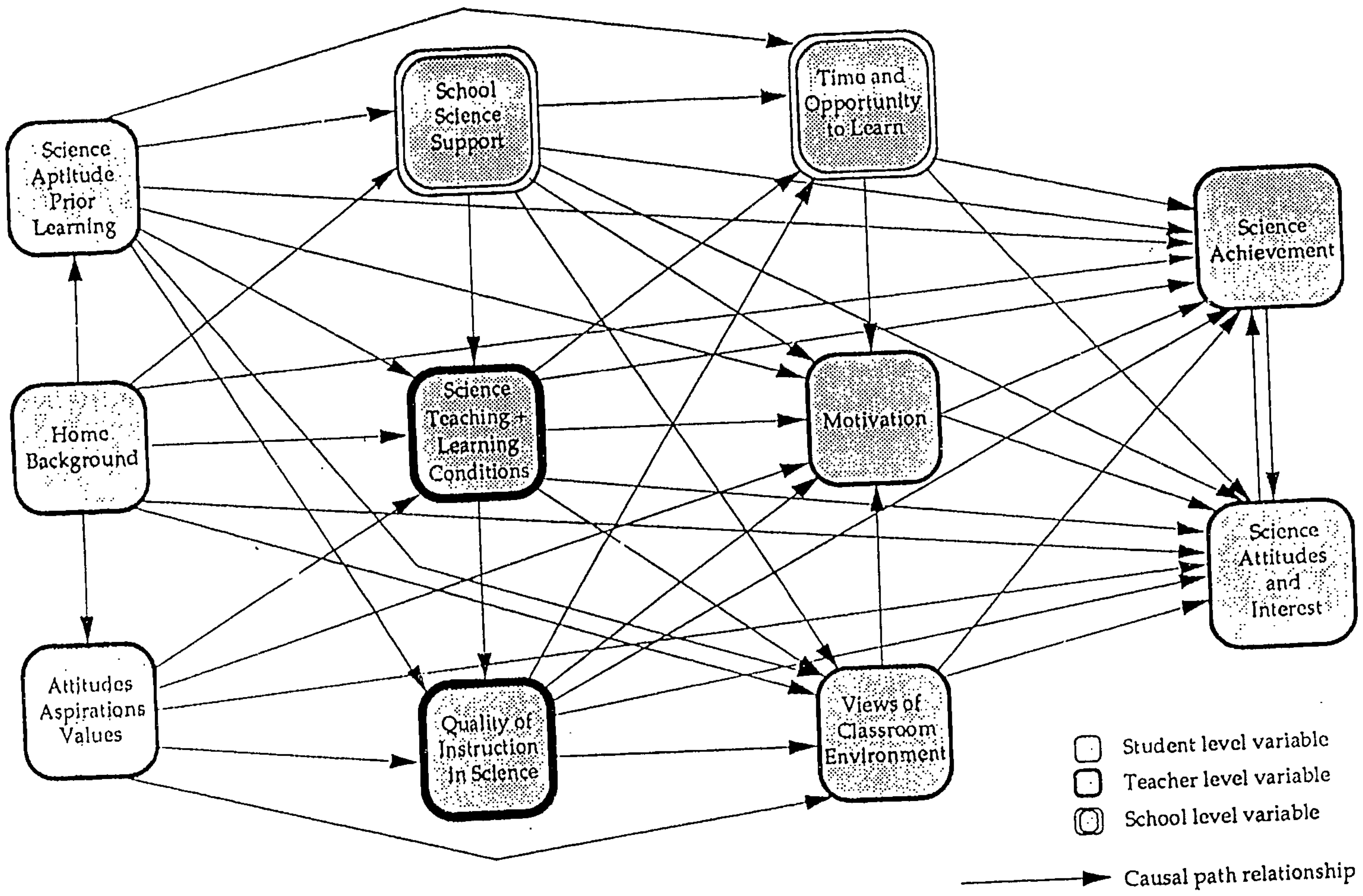
Science Aptitude Prior Learning
Home Background
Attitudes Aspirations Values
School Science Support
Science Teaching + Learning Conditions
Quality of Instruction In Science
Time and Opportunity to Learn
Motivation
Views of Classroom Environment
Science Achievement
Science Attitudes and Interest
Student level variable
Teacher level variable
School level variable
Causal path relationship

FIG. 2. A model of performance in science

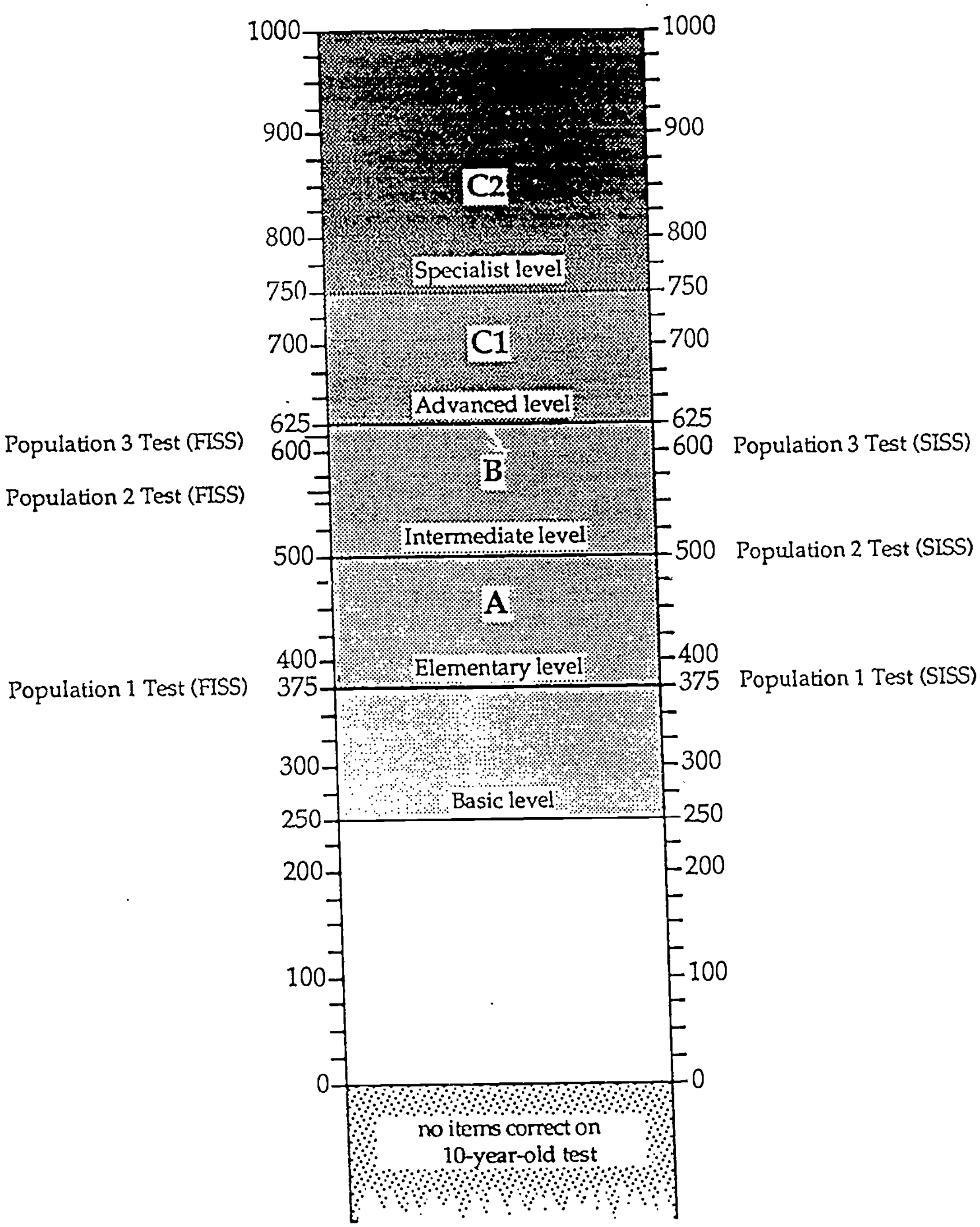

FIG. 3 The Science Achievement Scale

Basic Level 1A01 What gas in the air is essential for us to breathe in order to live?

A nitrogen
*B oxygen
C carbon dioxide
D hydrogen
E water vapour